Data Science

Maths

1)What is a vector?

> Quantity having both magnitude and direction.

2)How the addition of 2 vectors happens.

> Element-wise.

3)What will be the output of the below vector subtraction

a=[2,3],b=[18,4]

>) [-16,-1]

4)What is the concept behind the calculation of the length of the vector?

> Hypotenuse Theory

5)What is the length of the given vector?

f=[4,8]

> 8.94

6)Length of vector is what type of quantity

> Scaler

7)How many dimensions human eye can visualize

> 3D

8)What is the nature of linear equations?

a)If one quantity is increasing another is also increasing.

b)If one quantity is increasing another is decreasing

> Both a and b (correct)

9)What will be the output of AB.

$A = \begin{bmatrix} 1 & -2 & 1 \\ 2 & 1 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 1 & 1 \end{bmatrix}$

a) $\begin{bmatrix} -3 & -2 \\ 10 & 7 \end{bmatrix}$

b) $\begin{bmatrix} 4 & -3 & 5 \\ 7 & -4 & 9 \\ 3 & -1 & 4 \end{bmatrix}$

c)No output

d)Matrix multiplication is not possible.

> Correct answer is a

10)If a matrix has a row or a column with all elements equal to 0,then

determinant is

> 0

11)What will be the transpose of matrix

$4 = \begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix}$

a)Transpose is not possible
b)Transpose can be obtained by converting rows into columns and vice versa
c)By cross multiplication of elements
d)Transpose will result in null

> c

12)Two matrices are said to be equal only if the number of rows in both matrices should be

equal.

> True

13)Find the eigenvalues of the matrix

$$A = \begin{pmatrix} 8 & 0 & 0 \\ 6 & 6 & 11 \\ 1 & 0 & 1 \end{pmatrix}.$$

a)1 or 6 or 8

14)For matrix A, select the correct statement

>When A is squared, the eigen vectors remain same.The eigenvalues are squared

15)f(x,y)=4y3 + $2y$, $the\ partial\ derivative\ with\ respect\ to\ x\ is$

>0

==========

ML associate

1 . Linear regression is used to predict

> continuous target

2. The line with least SSR is called

> best fit line

3. Independent variables in linear regression can be

a. continuous

b. categorical

> both a and b (correct)

4. Regression line minimizes the

> Square of Residuals

5. For a linear regressor with three dependent variables, which equation can be used

> y = m0x0 + m1x1 + m2x2 + c

6. Which of the following is not an assumption of linear regression

> Residuals

7. Any anomalies in data will result in linear regression performing

> Bad

8. For some data, the predicted values are [23, 30, 33, 38] and the actual values are [25, 15, 40, 35]. Calculate the SSR.

> 287

9. When value of X increases, the value of Y decreases. X and Y are

> Negatively correlated

10. Linear regression is a form of

> Supervised ML

11. Logistic regression is used to predict

> categorical target

12. Logistic regression is used to predict

> Supervised ML

13. Logistic regression can do

> binary classification

14. Logistic Regression can use

A. continuous variables

B. discrete variables

C. non-linear features

D. all of the above (correct)

15. Which of the following is not a binary classification

A. spam / not spam

B. yes/no

C. positive / negative / recovering (correct)

D. true/false

16. Logistic Regression maps the probabilities to a

> Sigmoid Function

17. if sigmoid(X) = 0.9, the prediction is

> 1

18. The typical threshold between the 0 class and the 1 class in logistic regression is set at

> 0.5

19. Logistic Regression assumes linear relationship between the data

> FALSE

20. Logistic Regression needs independent variables to be normally distributed

> FALSE

21. Which of the following is not an application of Logistic regression

A. Spam filtering

B. Check if a part is defective or not

C. Check the stage of Diabetic Retinopathy (correct)

D. Check if a cell is cancerous or not

22. Which of the following evaluation metrics cannot be used for results of Logistic Regression

A. Accuracy Score

B. Confusion Matrix

C. Precision and Recall

D. r2 Score (correct)

23. Decision Tree is used to predict

A. continuous target

B. categorical target

C. both a and b (correct)

24. Decision Tree is a form of

> Supervised ML

25. Decision Tree performs better with

> categorical target

26. The topmost node in a decision tree is called

> Root Node

27. Nodes with no children are

> Leaf Node

28. Which of following is not a con of decision tree

A. Overfitting

B. Low accuracy on continuous variables

C. Small changes in data leading to large change in decision tree

D. None of the above (correct)

29. Decision trees make assumptions about the distribution of the data

> FALSE

30. Which of the following is a proof decision tree

A. Less data cleaning required

B. Can handle both continuous and categorical outputs

C. Easy to interpret

D. All of the above (correct)

****

ML associate

1. Which category of machine learning model do the learning methods of regression and classification fall under?

> semi-supervised learning

2. How is the predicted value calculated in a regression tree?

A. from the mode

B. from boosting

C. from the mean (correct)

D. from bagging

3. In a decision tree, what does each leaf represent?

a role or an option

A. an outcome

B. a timeline (correct)

C. a feature or an attribute

4. When is polynomial regression a strong predictive model choice to use?

A. when different models are built with different kernels

B. when there is a nonlinear relationship between the variables (correct)

C. when data scaling is not needed

D. when there is a linear relationship between the variables

5. When normalizing data, what are your values rescaled to?

A. a category

B. R squared

C. a zero to one range (correct)

D. any real number

6. Which function in pandas is used for one hot encoding?

> get_dummies()

7. Why is k-nearest neighbor also called lazy learning?

>It uses a lot of computation for every instance.

8. What characterizes an outlier in cluster analysis?

A. hopping from one centroid to another

B. being close to several centroids

C. falling exactly on one of the centroids

D. not being close to any centroid (correct)

9. Why are regression methods not considered good examples of machine learning?

A. They use plots instead of calculations.

B. They are based on algebraic models.

C. They are based on statistical predictions. (correct)

D. They use trend lines.

10. What should you do if your decision tree has too much entropy?

A. Delete the leaves.

B. Choose a different root.

C. Add more outcomes.

D. Add or substitute predictors. (correct)

11. Which one of the following do you use when you look for trends instead of trying to classify data into different groups?

A. binary classification

B. regression problems (correct)

C. multiclass classification

D. unsupervised learning

12. Salim wants to predict the best harvest date for his orchard based on prior weather reports and harvest histories. Which type of tool does he require?

A. multiclass classification

B. binary classification

C. regression analysis (correct)

D. transduction analysis

13. Which of the following is instance-based (or lazy learning)?

A. K-NN and Naïve Bayes

B. K-Means Clustering and K-NN (correct)

C. K-NN and Regression

D. K-Means Clustering and Regression

14. What distinguishes supervised machine learning from other types of machine learning?

A. identifying fundamental rules

B. classifying variables as dependent or independent

C. using programming or instruction

D. using labeled data for training (correct)

15. A company has a variety of efforts that it uses to find new antiviral medications. Which effort best exemplifies unsupervised learning?

A. the screening of a large collection of botanical extracts

****

1. The preProcess class can be used for many operations on predictors.
> True
Explanation: Operations include centering and scaling.
2. Which of the following function is used to generate the class distances?
> predict.classDist
Explanation: By default, the distances are logged.
3. Which of the following can also be used to find new variables that are linear combinations of the original set with independent components?
a) ICA
b) SCA
c) PCA
Answer: a
Explanation: ICA stands for independent component analysis.
4. The function preProcess estimates the required parameters for each operation.
> True
Explanation: predict.preProcess is used to apply them to specific data sets.
5. Which of the following can be used to impute data sets based only on information in the training set?
> preProcess
Explanation: This can be done with K-nearest neighbors.
6. Point out the correct statement.
> findLinearCombos will return a list that enumerates dependencies

Explanation: For each linear combination, it will incrementally remove columns from the matrix and test to see if the dependencies have been resolved.
7. The difference between the class centroids and the overall centroid is used to measure the variable influence
Explanation: The larger the difference between the class centroid and the overall center of the data, the larger the separation between the classes.
8. Which of the following model include a backward elimination feature selection routine?
a) MCV
b) MARS
c) MCRS
Answer: b
Explanation: MARS stands for Multivariate Adaptive Regression Splines.
9. The advantage of using a model-based approach is that is more closely tied to the model performance.
> True
Explanation: Model-based approach is able to incorporate the correlation structure between the predictors into the importance calculation.
10. Which of the following model sums the importance over each boosting iteration?
>Boosted trees
Explanation: gbm package can be used here.
11. Which of the following argument is used to set important values?
> scale
Explanation: All measures of importance are scaled to have a maximum value of 100.
12. For most classification models, each predictor will have separate variable importance for each class.
> True
Explanation: The exceptions are classification trees, bagged trees, and boosted trees.
13. Which of the following method can be used to combine different classifiers?
> Model stacking
Explanation: Model ensembling is also used for combining different classifiers.
14. Point out the correct statement.
c) Combining classifiers improve accuracy
Answer: c
Explanation: You can combine classifiers by averaging.
6. Which of the following function provides unsupervised prediction?
a) cl_forecast
b) cl_nowcast
c) cl_precast
d) none of the mentioned
Answer: d
Explanation: cl_predict function is clue package provides unsupervised prediction.

15. Model-based prediction considers relatively easy version for covariance matrix.
> False
Explanation: Model-based prediction considers relatively easy version for covariance matrix.

16. Which of the following is used to assist the quantitative trader in the development?
> quantmod
Explanation: Quandl package is similar to quantmod.
17. Which of the following function can be used for forecasting?
> forecast
Explanation: Forecasting is the process of making predictions of the future based on past and present data and analysis of trends.
18. Predictive analytics is same as forecasting.
> False
Explanation: Predictive analytics goes beyond forecasting
19. Which of the following is the correct formula for total variation?
> Total Variation = Residual Variation + Regression Variation
Explanation: The complementary part of the total variation is called unexplained or residual.
20. Point out the correct statement.
a) A standard error is needed to create a prediction interval
b) The prediction interval must incorporate the variability in the data around the line
c) Investors use the residual variance to measure the accuracy of their predictions on the value of an asset
d) All of the mentioned
Answer: d
Explanation: In statistics, explained variation measures the proportion to which a mathematical model accounts for the variation of a given data set.

21. Which of the following things can be accomplished with linear model?
a) Flexibly fit complicated functions
b) Uncover complex multivariate relationships
c) Build accurate prediction models
d) All of the mentioned
Answer: d
Explanation: Linear models are the single most important applied statistical and machine learning technique.

22. Which of the following statement is incorrect with respect to outliers?
>Outliers cannot conform to the regression relationship
Explanation: Outliers can conform to the regression relationship.
23. Point out the wrong statement.
a) The fraction of variance unexplained is an established concept in the context of linear regression
b) “Explained variance” is routinely used in principal component analysis
c) The general linear model extends simple linear regression (SLR) by adding terms linearly into the model
d) None of the mentioned
Answer: d
Explanation: Linearity refers to a mathematical relationship or function that can be graphically represented as a straight line.

24. Which of the following can be useful for diagnosing data entry errors?
hat values
Explanation: resid returns the ordinary residuals.
25. Multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and response.
> True
Explanation: Multivariate Data Analysis refers to any statistical technique used to analyze data that arises from more than one variable.
26. Residual _____ plots investigate normality of the errors.
a) RR
b) PP
c) QQ
Answer: c
Explanation: Patterns in your residual plots generally indicate some poor aspects of model fit.

27. Which of the following shows residuals divided by their standard deviations?
> rstandard
Explanation: rstandard stands for standardized residuals.
28. The least-squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationships.
> False
Explanation: Multivariate regression adjusts a coefficient for the linear impact of the other variables

Predicting with Regression

1. Predicting with trees evaluate _____ within each group of data.
> homogeneity
Explanation: Predicting with trees is easy to interpret.
2. Point out the wrong statement.
a) Training and testing data must be processed in different way
b) Test transformation would mostly be imperfect
c) The first goal is statistical and second is data compression in PCA
Answer: a
Explanation: Training and testing data must be processed in the same way.

3. Which of the following method options is provided by train function for bagging?
a) bagEarth
b) treebag
c) bagFDA
d) all of the mentioned
Answer: d
Explanation: Bagging can be done using bag function as well.

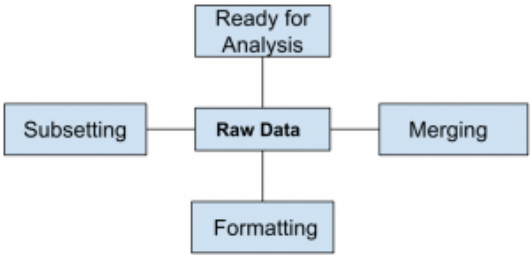
4. Which of the following is correct with respect to random forest?
a) Random forest are difficult to interpret but often very accurate

Answer: a
Explanation: Random forest is top performing algorithm in prediction.
5. Point out the correct statement.
a) Prediction with regression is easy to implement
b) Prediction with regression is easy to interpret
c) Prediction with regression performs well when linear model is correct
d) All of the mentioned
Answer: d
Explanation: Prediction with regression gives poor performance in non linear settings.

6. Which of the following library is used for boosting generalized additive models?
> gamBoost
Explanation: Boosting can be used with any subset of classifier.
7. The principal components are equal to left singular values if you first scale the variables.
> False
Explanation: The principal components are equal to left singular values if you first scale the variables.
8. Which of the following is statistical boosting based on additive logistic regression?
> gamBoost
Explanation: gamboost is used for model based boosting.
9. Which of the following is one of the largest boost subclass in boosting?
> gradient boosting
Explanation: R has multiple boosting libraries.
10. PCA is most useful for non linear type models.
> False
Explanation: PCA is most useful for linear type models.

Raw and Processed Data

1. Which of the following block information is odd man out?



> Raw data
Explanation: Characteristics mentioned in the diagram are traits of processed data.
2. Point out the correct statement.
>Data has only qualitative value
Explanation: Data belongs to the set of items.
3. Data that summarize all observations in a category are called _____ data.
> summarized

Explanation: The summary could be the sum of the observations, the number of occurrences, their mean value, and so on.

4. Which of the following is an example of raw data?

a) original swath files generated from a sonar system

b) initial time-series file of temperature values

c) a real-time GPS-encoded navigation file

d) all of the mentioned

Answer: d

Explanation: Raw data refers to data that have not been changed since acquisition.

5. Point out the correct statement.

>Primary data is original source of data

Explanation: Primary data is also referred to as raw data.

6. Which of the following data is put into a formula to produce commonly accepted results?

> Processed

Explanation: Raw data came from direct measurements.

7. Processing data includes subsetting, formatting and merging only.

> False

Explanation: There are many other techniques applied to raw data.

8. Which of the following is another name for raw data?

> eggy data

Answer: b

Explanation: Although raw data has the potential to become “information,” extraction, organization, and sometimes analysis and formatting for presentation are required for that to occur.

9. Which type of data is generated by POS terminal in a busy supermarket each day?

a) Source

b) Processed

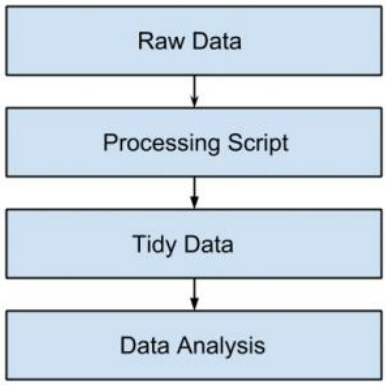
c) Synchronized

d) All of the mentioned

Answer: a

Explanation: Raw data is sometimes referred to as source data.

10. Following figure represents correct sequence of steps in performing data analysis.



> True

Explanation: Data analysis is not a goal in itself; the goal is to enable the business to make better decisions.

Summarizing and Merging Data

1. Which of the following function gives information about top level data?
> head
Explanation: The function head is very useful for working with lists, tables, data frames and even functions.
2. Point out the correct statement.
a) head function work on string
b) tail function work on string
c) head function work on string but tail function do not
d) none of the mentioned
Answer: d
Explanation: Both head and tail function do not work on strings.

3. Which of the following function is used for quantiles of quantitative values?
> quantile

Explanation: In probability and statistics, the quantile function specifies, for a given probability in the probability distribution of a random variable, the value at which the probability of the random variable will be less than or equal to that probability.
4. Which of the following function is used for determining missing values?
a) any
b) all
c) is
d) all of the mentioned
Answer: d
Explanation: In R, missing values are represented by the symbol NA.

5. Point out the wrong statement.
a) Common variables are used to create missingness vector
b) Common variables are used to cutting up quantitative variables
c) Common variables are not used to apply transforms
Answer: c
Explanation: Common variables are not used to apply transforms.
6. Which of the following transforms can be performed with data value?
a) log2
b) cos
c) log10
d) all of the mentioned
Answer: d
Explanation: Many common transforms can be applied to the data with R.
7. Each observation forms a column in tidy data.
> False
Explanation: Each variable forms a column in tidy data.

8. Which of the following function is used for casting data frames?
> dcast
Explanation: Use acast or dcast depending on whether you want vector/matrix/array output or data frame output.
9. Which of the following join is by default used in plyr package?
>left
Explanation: Join is faster in plyr package.
10. mutate function is used for casting as multi dimensional arrays.
>False
Explanation: mutate is used for adding new variables.
5. Point out the wrong statement.
a) ROC curve stands for receiver operating characteristic
b) Foretime series, data must be in chunks
c) Random sampling must be done with replacement
d) None of the mentioned
Answer: d
Explanation: Random sampling with replacement is the bootstrap.

6. Which of the following is a categorical outcome?
a) RMSE
b) RSquared
c) Accuracy
Answer: c
Explanation: RMSE stands for Root Mean Squared Error.

7. For k cross-validation, larger k value implies more bias.
>False
Explanation: For k cross-validation, larger k value implies less bias.
8. Which of the following method is used for trainControl resampling?
> repeatedcv
Explanation: repeatedcv stands for repeated cross-validation.
9. Which of the following can be used to create the most common graph types?
> qplot
Explanation: qplot() is short for a quick plot.
10. For k cross-validation, smaller k value implies less variance.
> True
Explanation: Larger k value implies more variance.

1 . Which one among these are immutable?

> Tuple

2. python packages are namespaces containing _____.

> multiple modules

3. What executes the function given as the first argument on all the elements of the iterable given as the

second argument.

> map function

4. In which data structure, the elements are arranged in ascending order?

> Sets

5. Array elements can be removed using_____ method.

A. pop()

B. remove()

D. both a and b

6. Syntax for displaying last before element of size 6 is

D. arr[-2]

7. Which one is called anonymous function?

> lambda

8. arr.pop() is used to delete

> last element is removed

9. str = "India recovering from Covid" Print alternate characters(includes space in character)

> str[::2]

10. Does python have switch case statement?

> FALSE

11. Pass in Python indicates

> no operation

12. The method that returns a new set with items common to two sets is

A. union()

B. issubset()

C. intersection() (correct)

D. isdisjoint()

13. consider the string str="Certified Data Scientist" , Extract Data from this string

> str[10:14]

14. Print DataDataData by using the string str="Certified Data Scientist"

> str[10:14]*3

15. What will be the output of the following code?

```
p = 7
q = 9
print ((p > q) | (p == q))
print ((p > q) & (p == q))
print ((p < q) | (p > q))
print ((p < q) & (p > q))
```

> False,False,True,False

16. str = "Hai Everyone! How are you all?" Print alternate characters(includes space in character)

```
> str[::2]
```

1.Which of the following are correct component for data science?

A. Data Engineering

B. Advanced Computing

C. Domain expertise

D. All of the above

2. Data Science can be simply stated as

> Insights from Data

3. Which of the following allows you to find the relationship you didn't about?

A. Inferential

B. Exploratory

C. Causal

D. None of the mentioned

4. Which of the following term is appropriate to volume, variety, velocity?

> Big data

5. Data Science uses

> structured and unstructured data

6. Which uses labelled (past) data to learn and predict future events

> supervised ML

7. Point out the correct statement.

> Raw data is original source of data

8. Which of the following is performed by Data Scientist?

A. Define the question

B. Create reproducible code

C. Challenge results

D. All of the mentioned (correct)

9. Which of the following is not a application for data science?

A. Recommendation Systems

B. Image & Speech Recognition

C. Privacy Checker(correct)

D. Online Price Comparison

10. Raw data should be processed only one time.

> FALSE

11. Which of the following approach should be used to ask Data Analysis question?

> Find out the question which is to be answered

12. Data Science is a subset of

A. ML, DL

B. AI

C. AI,ML,DL

D. None of the Above(correct)

13. Point out the correct statement.

A. Descriptive analysis is first kind of data analysis performed

B. Descriptions can be generalized without statistical modelling.(correct)

C. Description and Interpretation are same in descriptive analysis

14. Which of the following language is used in Data science?

> R

15. Which of the following is false?

A. Subsetting can be used to select and exclude variables and observations

B. Raw data should be processed only one time.(correct)

C. Merging concerns combining datasets on the same observations to produce a result with more variables

16. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics

B. Machine Learning / Deep Learning

C. Data Wrangling

D. All of the above (correct)

17. Which of the following is not a part of data science process?

A. Discovery

B. Model Planning

C. Communication Building (correct)

D. Operationalize

18. 20% of the work in Data science constitutes

A. Preprocessing

B. modelling, Evaluation (correct answer)

C. EDA

D. Domain checking

19. The Data Science Venn Diagram is composed of Hacking Skills, Math or Statistics Skill and---

A. Domain Expertise(correct)

B. Probability

C. Optimization

D. None of the mentioned

20. Which of the following is the common goal of statistical modelling?

A. Inference (correct)

B. Summarizing

C. Subsetting

D. None of the mentioned

Pandas

1 . In pandas, Index values must be?

A. unique

2. PANDAS stands for _____

C. Panel Data

3. In pandas, dataframe contains

> 2D data

4. _____ is used when data is in Tabular Format.

> pandas

5. A _____ is a one-dimensional array.

> Series

6. Which of the following operation works with the same syntax as the analogous dict operations?

A. Getting columns

B. Setting columns

C. Deleting columns

D. All of the above (correct)

7. Which of the following input can be accepted by DataFrame?

A. Structured ndarray

B. Series

C. DataFrame

D. All of the mentioned (correct)

8. A Series by default have numeric data labels starting from

> 0

9. Which of the following object you get after reading CSV file?

> DataFrame

10. pd.read_csv() function can be used to read files with the extension .txt

> TRUE

11. If the missing values are removed from a column containing values of integer data type, the data type of that column automatically becomes an integer

> TRUE

12. Which of the following commands is used to remove the rows of df where at least one column has NaN values?

> df.dropna(how='any', inplace=True)

13. Considering the dataframe df.isnull() will give you ----------result

> Boolean

14. NaN acts as a place holder for missing values

> TRUE

15. used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.

> describe()

16. Mention two methods for filling the nan values in a dataframe

> df.fillna(method=bfill), cars.mean()

17. Series is a one-dimensional labeled array capable of holding any data type.

> TRUE

****

Numpy

1 . Which of the following is contained in NumPy library?

a) n-dimensional array object

b) tools for integrating C/C++ and Fortran code

c) fourier transform

d) all of the mentioned

2. To create sequences of numbers, NumPy provides a function _____ analogous to range that returns arrays instead of lists.

> arange

3. Point out the correct statement.

a) NumPy main object is the homogeneous multidimensional array

b) In Numpy, dimensions are called axes

c) Numpy array class is called ndarray

d) All of the mentioned

4. Which of the following method creates a new array object that looks at the same data?

> copy

5. Which of the following attribute should be used while checking for type combination input and output?

> types

6. Which of the following returns an array of ones with the same shape and type as a given array?

> ones

7. What is/are the advantage(s) of NumPy Arrays over classic Python lists?

A. Insertion, concatenation and deletion are faster.

B. It has better support for mathematical operations.

C. They consume less memory.

D. All of these answers.(correct)

8. The most important object defined in NumPy is an N-dimensional array type called?

> ndarray

9. What will be output for the following code?

A. A. [[1, 2, 3]]

B. B. [1]

C. C. [1, 2, 3]

D. D. Error

10. Which of the following statement is false?

A. ndarray is also known as the axis array.(correct)

B. ndarray.dataitemSize is the buffer containing the actual elements of the array.

C. NumPy main object is the homogeneous multidimensional array

D. In Numpy, dimensions are called axes

11. If a dimension is given as ____ in a reshaping operation, the other dimensions are automatically calculated.

> Negative one

12. What will be printed?

A. 7

B. 8

C. 10

D. 21

13. import numpy as np

14. a = np.array([1,2,3,5,8])

15. b = np.array([0,3,4,2,1])

16. c = a + b

17. c = c*a

18. print (c[2])

19. What will be output for the following code?

A. 1

B. 4

C. 5

D. 6

20. import numpy as np

21. a = np.array([[1,2,3],[0,1,4]])

22. print (a.size)

23. reshape() function in numpy array using python is:

> reshape(array, shape)

24. How to convert numpy array to list?

> list(array)

25. What is the correct code to install numpy in the windows

system containing python3?

> pip install numpy

Statistics

1 . Which of the following is a branch of statistics?

A. Descriptive statistics

B. Inferential statistics

C. Both A and B

2. Find the median of the call received on 7 consecutive days 11,13, 17, 13, 23,25,19

> 17

3. When the Mean of a number is 21, what is the Mean of the sampling distribution?

> 21

4. Find the variance of the given data set: 3,9,5,6,7

> 4

5. Which of the following values is used as a summary measure for a sample, such as a sample mean?

> Sample statistic

6. What are the variables whose calculation is done according to the weight, height, and length known as?

> Continuous variables

7. What is the scale applied in statistics, which imparts a difference of magnitude and

proportions, is considered as?

> Ratio scale

8. What is true about Statistics?

> Statistics is used to process complex problems in the real world

9. SciPy stands for?

> Scientific Python library

10. An observation point that is distant from other observations

> outliers

11. Which is a measure of the width of a distribution

> Variance

12. Which is completed characterized by mean and variance

> Normal Distribution

13. The parametric statistical tests, like t-test and ANOVA assumes

> normally-distributed data

14. Which is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean

> skewness

15. Point out the wrong statement.

A. A random variable is a numerical outcome of an experiment

B. Continuous random variable can take any value on the real line

C. There are three types of random variable (correct)

16. Which test can be performance on two or more samples of quantitative data ?

A. One-way ANOVA (correct)

B. p-test

C. t-test

D. chi-square test

17. The major categories of data are

> numerical & categorical

Statistics

1. What is the Central Limit Theorem and why is it important?

“Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.” Read more here.

2. What is sampling? How many sampling methods do you know?

“Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.” Read the full answer here.

3. What is the difference between type I vs type II error?

“A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.” Read the full answer here.

4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?

A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship. Read more here and here.

5. What are the assumptions required for linear regression?

There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

6. What is a statistical interaction?

”Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor.” *Read more* here.

7. What is selection bias?

“Selection (or ‘sampling’) bias occurs in an ‘active,’ sense when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases ..

the model will see. That is, active selection bias occurs when a subset of the data are systematically (i.e., non-randomly) excluded from analysis." *Read more* here.

8. What is an example of a data set with a non-Gaussian distribution?

"The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate." *Read more* here.

9. What is the Binomial Probability Formula?

"The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring."

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years? The answer lies in the difference between explaining and predicting.

The differences between supervised and unsupervised learning are as follows;

Input data is labelled.	Input data is unlabelled.
Uses a training data set.	Uses the input data set.
Used for prediction.	Used for analysis.
Enables classification and regression.	Enables Classification, Density Estimation, & Dimension Reductio

****

Q1: Mention three ways to make your model robust to outliers.

1. Investigating the outliers is always the first step in understanding how to treat them. After you understand the nature of why the outliers occurred you can apply one of the several methods mentioned below.
2. Add regularization that will reduce variance, for example, L1 or L2 regularization.
3. Use tree-based models (random forest, gradient boosting) that are generally less affected by outliers.
4. Winsorize the data. Winsorizing or winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. In numerical data, if the distribution is almost normal using the Z-score we can detect the outliers and treat them by either removing or capping them with some value. If the distribution is skewed using IQR we can detect and treat it by again either removing or capping it with some value. In categorical data check for value_count in the percentage if we have very few records from some category, either we can remove it or can cap it with some categorical value like others.
5. Transform the data, for example, you do a log transformation when the response variable follows an exponential distribution or is right-skewed.
6. Use more robust error metrics such as MAE or Huber loss instead of MSE.

7. Remove the outliers, only do this if you are certain that the outliers are true anomalies that are not worth adding to your model. This should be your last consideration since dropping them means losing information.

Q2: Describe the motivation behind random forests and mention two reasons why they are better than individual decision trees.

The motivation behind random forest or ensemble models in general in layman's terms, Let's say we have a question/problem to solve we bring 100 people and ask each of them the question/problem and record their solution. Next, we prepare a solution which is a combination/ a mixture of all the solutions provided by these 100 people. We will find that the aggregated solution will be close to the actual solution. This is known as the "Wisdom of the crowd" and this is the motivation behind Random Forests. We take weak learners (ML models) specifically, Decision Trees in the case of Random Forest & aggregate their results to get good predictions by removing dependency on a particular set of features. In regression, we take the mean and for Classification, we take the majority vote of the classifiers.

A random forest is generally better than a decision tree, however, you should note that no algorithm is better than the other it will always depend on the use case & the dataset Check the No Free Lunch Theorem. Reasons why random forests allow for stronger prediction than individual decision trees:

8. Decision trees are prone to overfit whereas random forest generalizes better on unseen data as it is using randomness in feature selection as well as during sampling of the data. Therefore, random forests have lower variance compared to that of the decision tree without substantially increasing the error due to bias.
9. Generally, ensemble models like Random Forest perform better as they are aggregations of various models (Decision Trees in the case of Random Forest), using the concept of the "Wisdom of the crowd."

Q3: What are the differences and similarities between gradient boosting and random forest? and what are the advantage and disadvantages of each when compared to each other?

Similarities:

10. Both these algorithms are decision-tree-based algorithms
11. Both these algorithms are ensemble algorithms
12. Both are flexible models and do not need much data preprocessing.

Differences:

13. Random forests (Uses Bagging): Trees are arranged in a parallel fashion where the results of all trees are aggregated at the end through averaging or majority vote.
    Gradient boosting (Uses Boosting): Trees are arranged in a series sequential fashion where every tree tries to minimize the error of the previous trees.
14. Radnom forests: Every tree is constructed independently of the other trees.
    Gradient boosting: Every tree is dependent on the previous tree.

Advantages of gradient boosting over random forests:

15. Gradient boosting can be more accurate than Random forests because we train them to minimize the previous tree's error.
16. Gradient boosting is capable of capturing complex patterns in the data.
17. Gradient boosting is better than random forest when used on unbalanced data sets.

Advantages of random forests over gradient boosting :

18. Radnom forest is less prone to overfit as compared to gradient boosting.
19. The random forest has faster training as trees are created parallelly & independently of each other.

The disadvantage of GB over RF:

20. Gradient boosting is more prone to overfitting than random forests due to their focus on mistakes during training iterations and the lack of independence in tree building.
21. If the data is noisy the boosted trees might overfit and start modeling the noise.
22. In GB training might take longer because every tree is created sequentially.
23. Tunning the hyperparameters of gradient boosting is harder than those of random forest.

Q4: What are L1 and L2 regularization? What are the differences between the two?

Answer:

Regularization is a technique used to avoid overfitting by trying to make the model more simple. One way to apply regularization is by adding the weights to the loss function. This is done in order to consider minimizing unimportant weights. In L1 regularization we add the sum of the absolute of the weights to the loss function. In L2 regularization we add the sum of the squares of the weights to the loss function.

So both L1 and L2 regularization are ways to reduce overfitting, but to understand the difference it's better to know how they are calculated:

Loss (L2): Cost function + L * weights ^{2}

Loss (L1) : Cost function + L * |weights|

Where L is the regularization parameter

1- L2 regularization penalizes huge parameters preventing any of the single parameters to get too large. But weights never become zeros. It adds parameters square to the loss. Preventing the model from overfitting any single feature.

2 — L1 regularization penalizes weights by adding a term to the loss function which is the absolute value of the loss. This leads to it removing small values of the parameters leading in the end to the parameter hitting zero and staying there for the rest of the epochs. Removing this specific variable completely from our calculation. So, It helps in simplifying our model. It is also helpful for feature selection as it shrinks the coefficient to zero which is not significant in the model.

Q5: What are the Bias and Variance in a Machine Learning Model and explain the bias-variance trade-off?

Answer:

The goal of any supervised machine learning model is to estimate the mapping function (f) that predicts the target variable (y) given input (x). The prediction error can be broken down into three parts:

Bias: The bias is the simplifying assumption made by the model to make the target function easy to learn. Low bias suggests fewer assumptions made about the form of the target function. High bias suggests more assumptions made about the form of the target data. The smaller the bias error the better the model is. If the bias error is high, this means that the model is underfitting the training data.

Variance: Variance is the amount that the estimate of the target function will change if different training data was used. The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables. If the variance error is high this indicates that the model overfits the training data.

Irreducible error: It is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable. The irreducible error cannot be reduced regardless of what algorithm is used.

The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn, the algorithm should achieve good prediction performance. The parameterization of machine learning algorithms is often a battle to balance out bias and variance. For example, if you want to predict the housing prices given a large set of potential predictors. A model with high bias but low variance, such as linear regression will be easy to implement, but it will oversimplify the problem resulting in high bias and low variance. This high bias and low variance would mean in this context that the predicted house prices are frequently off from the market value, but the value of the variance of these predicted prices is low. On the other side, a model with low bias and high variance such as a neural network will lead to predicted house prices closer to the market value, but with predictions varying widely based on the input features.

Q6: Mention three ways to handle missing or corrupted data in a dataset?

Answer:

In general, real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values. However, you should start by asking the data owner/stakeholder about the missing or corrupted data. It might be at the data entry level, because of file encoding, etc. which if aligned, can be handled without the need to use advanced techniques.

There are different ways to handle missing data, we will discuss only three of them:

24. **Deleting the row with missing values**

The first method to handle missing values is to delete the rows or columns that have null values. This is an easy and fast method and leads to a robust model, however, it will lead to the loss of a lot of information depending on the amount of missing data and can only be applied if the missing data represent a small percentage of the whole dataset.

2. Using learning algorithms that support missing values

Some machine learning algorithms are robust to missing values in the dataset. The K-NN algorithm can ignore a column from a distance measure when there are missing values. Naive Bayes can also support missing values when making a prediction. Another algorithm that can handle a dataset with missing values or null values is the random forest model and Xgboost (check the post in the first comment), as it can work on non-linear and categorical data. The problem with this method is that these models' implementation in the scikit-learn library does not support handling missing values, so you will have to implement it yourself.

3. Missing value imputation

Data imputation means the substitution of estimated values for missing or inconsistent data in your dataset. There are different ways to estimate the values that will replace the missing value. The simplest one is to replace the missing value with the most repeated value in the row or the column. Another simple way is to replace it with the mean, median, or mode of the rest of the row or the column. This advantage of this is that it is an easy and fast way to handle the missing data, but it might lead to data leakage and does not factor in the covariance between features. A better way is to use a machine learning model to learn the pattern between the data and predict the missing values, this is a very good method to estimate the missing values that will not lead to data leakage and will factor the covariance between the feature, the drawback of this method is the computational complexity especially if your dataset is large.

Q7: Explain briefly the logistic regression model and state an example of when you have used it recently?

Answer:

Logistic regression is used to calculate the probability of occurrence of an event in the form of a dependent output variable based on independent input variables. Logistic regression is commonly used to estimate the probability that an instance belongs to a particular class. If the probability is bigger than 0.5 then it will belong to that class (positive) and if it is below 0.5 it will belong to the other class. This will make it a binary classifier.

It is important to remember that the Logistic regression isn't a classification model, it's an ordinary type of regression algorithm, and it was developed and used before machine learning, but it can be used in classification when we put a threshold to determine specific categories"

There are a lot of classification applications to it:

Classify email as spam or not, To identify whether the patient is healthy or not, and so on.

Q8:Explain briefly batch gradient descent, stochastic gradient descent, and mini-batch gradient descent? and what are the pros and cons for each of them?

Gradient descent is a generic optimization algorithm cable for finding optimal solutions to a wide range of problems. The general idea of gradient descent is to tweak parameters iteratively in order to minimize a cost function.

1.Batch Gradient Descent: In Batch Gradient descent the whole training data is used to minimize the loss function by taking a step towards the nearest minimum by calculating the gradient (the direction of descent)

Pros:

Since the whole data set is used to calculate the gradient it will be stable and reach the minimum of the cost function without bouncing (if the learning rate is chosen correctly)

Cons:

Since batch gradient descent uses all the training sets to compute the gradient at every step, it will be very slow especially if the size of the training data is large.

2. Stochastic Gradient Descent: Stochastic Gradient Descent picks up a random instance in the training data set at every step and computes the gradient-based only on that single instance.

Pros:

25. It makes the training much faster as it only works on one instance at a time.
26. It become easier to train large datasets

Cons:

Due to the stochastic (random) nature of this algorithm, this algorithm is much less regular than the batch gradient descent. Instead of gently decreasing until it reaches the minimum, the cost function will bounce up and down, decreasing only on average. Over time it will end up very close to the minimum, but once it gets there it will continue to bounce around, not settling down there. So once the algorithm stops the final parameter are good but not optimal. For this reason, it is important to use a training schedule to overcome this randomness.

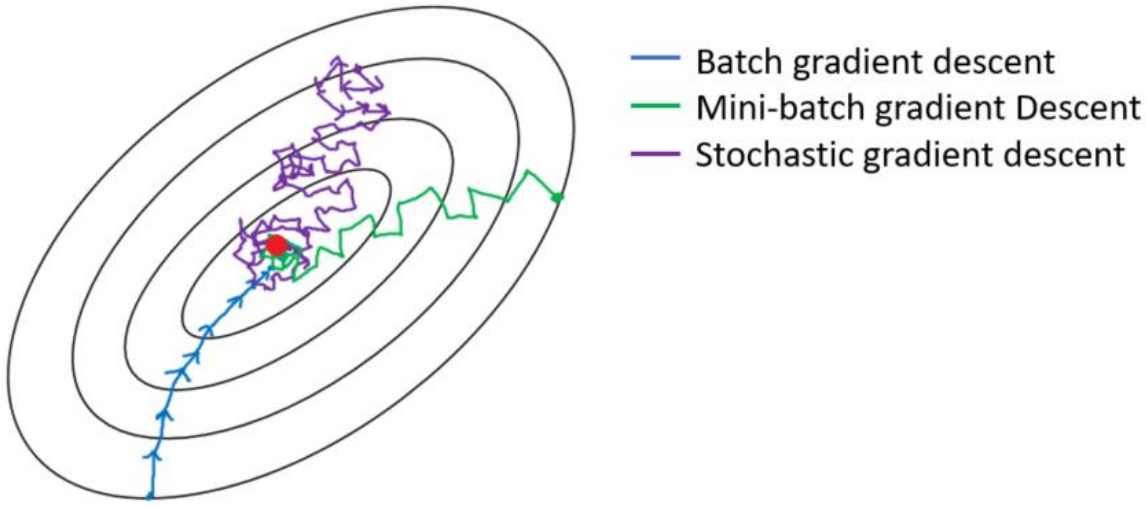
3. Mini-batch Gradient: At each step instead of computing the gradients on the whole data set as in the Batch Gradient Descent or using one random instance as in the Stochastic Gradient Descent, this algorithm computes the gradients on small random sets of instances called mini-batches.

Pros:

27. The algorithm's progress space is less erratic than with Stochastic Gradient Descent, especially with large mini-batches.
28. You can get a performance boost from hardware optimization of matrix operations, especially when using GPUs.

Cons:

29. It might be difficult to escape from local minima.



The difference between batch gradient descent, mini-batch gradient descent, and stochastic gradient descent.

Q9: Explain what is information gain and entropy in the context of decision trees?

Entropy and Information Gain are two key metrics used in determining the relevance of decision making when constructing a decision tree model and to determine the nodes and the best way to split.

The idea of a decision tree is to divide the data set into smaller data sets based on the descriptive features until we reach a small enough set that contains data points that fall under one label.

Entropy is the measure of impurity, disorder, or uncertainty in a bunch of examples. Entropy controls how a Decision Tree decides to split the data. Information gain calculates the reduction in entropy or surprise from transforming a dataset in some way. It is commonly used in the construction of decision trees from a training dataset, by evaluating the information gain for each variable, and selecting the variable that maximizes the information gain, which in turn minimizes the entropy and best splits the dataset into groups for effective classification.

Q10: Explain the linear regression model and discuss its assumption?

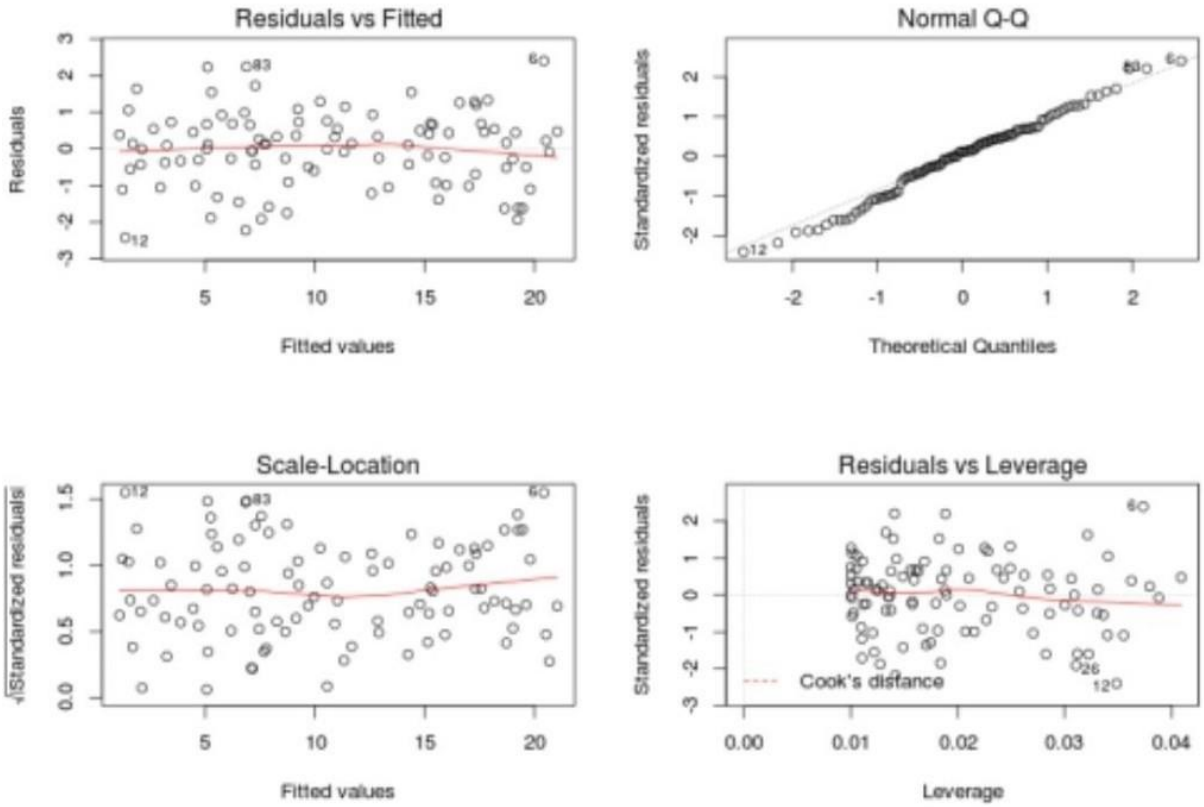
Linear regression is a supervised statistical model to predict dependent variable quantity based on independent variables. Linear regression is a parametric model and the objective of linear regression is that it has to learn coefficients using the training data and predict the target value given only independent values.

Some of the linear regression assumptions and how to validate them:

30. A linear relationship between independent and dependent variables
31. Independent residuals and the constant residuals at every x We can check for 1 and 2 by plotting the residuals(error terms) against the fitted values (upper left graph). Generally, we should look for a lack of patterns and a consistent variance across the horizontal line.
32. Normally distributed residuals We can check for this using a couple of methods:
    - Q-Q-plot(upper right graph): If data is normally distributed, points should roughly align with the 45-degree line.
    - Boxplot: it also helps visualize outliers
    - Shapiro–Wilk test: If the p-value is lower than the chosen threshold, then the null hypothesis (Data is normally distributed) is rejected.

4. Low multicollinearity
    - you can calculate the VIF (Variable Inflation Factors) using your favorite statistical tool. If the value for each covariate is lower than 10 (some say 5), you're good to go.

The figure below summarizes these assumptions.

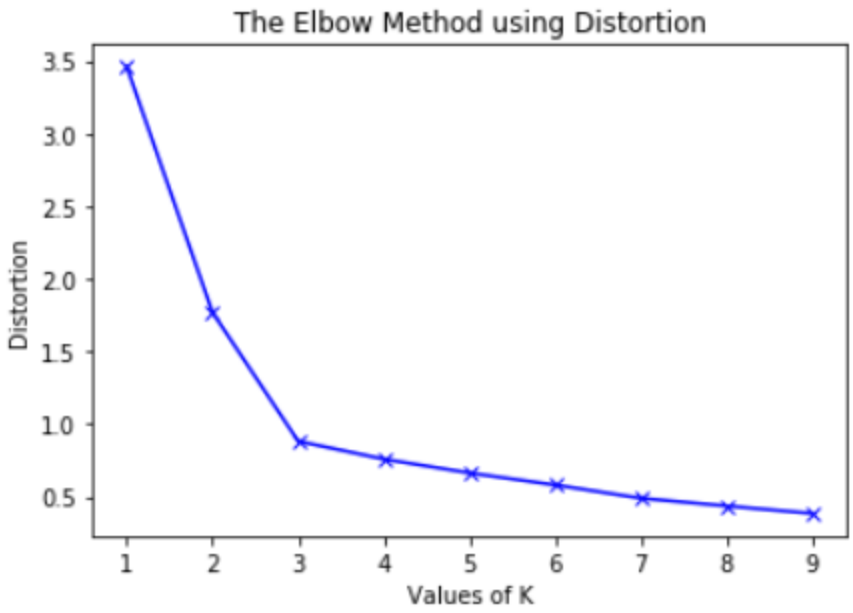


The linear regression assumption.

Q11: Explain briefly the K-Means clustering and how can we find the best value of K?

K-Means is a well-known clustering algorithm. K-Means clustering is often used because it is easy to interpret and implement. It starts by partitioning a set of data into K distinct clusters and then arbitrary selects centroids of each of these clusters. It iteratively updates partitions by first assigning the points to the closet cluster and then updating the centroid and then repeating this process until convergence.

The process essentially minimizes the total inter-cluster variation across all clusters. The elbow method is a well-known method to find the best value of K in K-means clustering. The intuition behind this technique is that the first few clusters will explain a lot of the variation in the data, but past a certain point, the amount of information added is diminishing. Looking at the graph below of the explained variation (on the y-axis) versus the number of cluster K (on the x-axis), there should be a sharp change in the y-axis at some level of K. For example in the graph below the drop-off is at k=3.



The elbow method to find the best value of K.

The explained variation is quantified by the within-cluster sum of squared errors. To calculate this error notice, we look for each cluster at the total sum of squared errors using Euclidean distance.

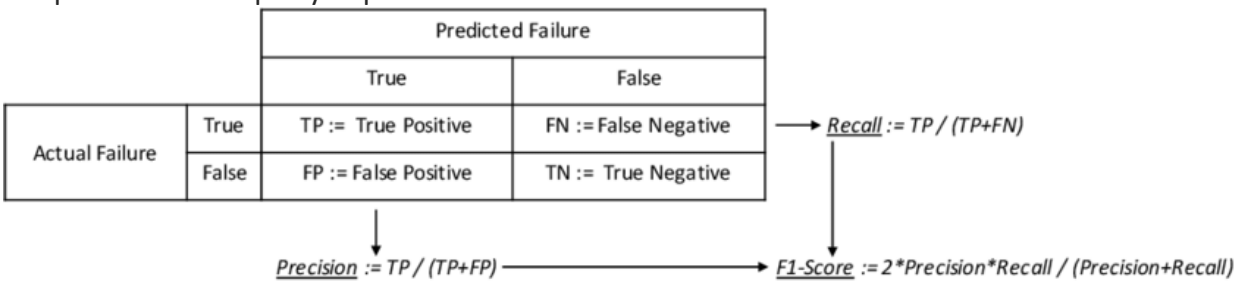
Another popular alternative method to find the value of K is to apply the silhouette method, which aims to measure how similar points are in its cluster compared to other clusters. It can be calculated with this equation: (x-y)/max(x,y), where x is the mean distance to the examples of the nearest cluster, and y is the mean distance to other examples in the same cluster. The coefficient varies between -1 and 1 for any given point. A value of 1 implies that the point is in the right cluster and the value of -1 implies that it is in the wrong cluster. By plotting the silhouette coefficient on the y-axis versus each K we can get an idea of the optimal number of clusters. However, it is worthy to note that this method is more computationally expensive than the previous one.

Q12: Define Precision, recall, and F1 and discuss the trade-off between them?

Precision and recall are two classification evaluation metrics that are used beyond accuracy. Consider a classification task with two classes. **Precision** is the actual positive positive proportion of observations that were predicted positive by the classifier. **Recall** is the percentage of total positive cases captured out of all positive cases.

In the real world, there is always a trade-off between optimizing for precision and recall. Consider you are working on a classification task for classifying cancer patients from healthy people. Optimizing the model to have only high recall will mean that the model will catch most of the people with cancer but at the same time, the number of misdiagnosed people with cancer will increase. This will subject healthy people to dangerous and costly cancer treatments. On the other hand, optimizing the model to have high precision will make the model confident about the diagnosis, in favor of missing some people who truly have the disease. This will lead to fatal outcomes as they will not be treated. Therefore it is important to

optimize both precision and recall and the percentage of importance of each of them will depend on the application you are working on.
F1 score is the harmonic mean of precision and recall, and it is calculated using the following formula: F1 = 2*(precision*recall) / (precision+recall). The F1 score is used when the recall and the precision are equally important.



The precision, recall, and F1 score

****

Visualization

1 . matplotlib.pyplot by convention is imported as

> plt

2. plt.plot will draw which kind of plot

> line

3. Histogram is a plot of

> Frequency

4. Appropriate bin size for histogram for all possible data is

> Based on the insight being explored

5. Which of the following plots cannot be used to check coorelation between two predictors

A. Pie Chart (correct answer)

B. Line Chart

C. Scatter Plot

D. Bar Chart

6. Seaborn is based on which of the following libraries?

> matplotlib

7. Which method of the following can be used to plot a histogram using matplotlib

A. .hist (correct)

B. .distplot

C. .bar

D. .line

8. df.plot() will (assuming df is a DataFrame)

> Plot the DataFrame data

9. To create a scatter plot which method of pyplot can be used

>scatter

10. Suppose you want to visualize which quarter of year resulted in what ratio of yearly earnings. The appropriate plot for this

> Pie Chart

11. To add title to a plot, we can use which of the following methods

>title

12. You have age vs number of icecreams sold as data given to you and suppose you to visualize how the icecream is selling in various age groups, the appropriate plot for this will be

> Histogram

Pandas

1.What is a correct syntax to create a Pandas Series from a Python list?

> pd.Series(mylist)

2.What is a correct syntax to return the first value of a Pandas Series?

> myseries[0]

3.What is a correct syntax to add the lables "x", "y", and "z" to a Pandas Series?

```
> pd.Series(mylist, index = ["x", "y", "z"])
```

4.What is a correct syntax to create a Pandas DataFrame?

```
> pd.DataFrame(data)
```

5.What is a correct syntax to return the first row in a Pandas DataFrame?

> Df.loc[0]

6.What is a correct syntax to return both the first row and the second row in a Pandas DataFrame?
> Df.loc[[0,1]]

7. What is the correct Pandas function for loading CSV files into a DataFrame?

> read_csv()

8. What is a correct syntax to return the first 20 rows of a DataFrame?

> df.head(20)

9. What is a correct syntax to return the entire DataFrame?

> df.to_string()

10. What is the correct Pandas function for loading JSON files into a DataFrame?

> read_json()

11. What is a correct syntax to load a Python Dictionary called "data" into a Pandas DataFrame?

> pd.DataFrame(data)

12. What does the Pandas head() method do?

> Returns the headers and the specified number of rows, starting from the top

13. For the Pandas head() method, how many rows are returned by default, if you do not specify it?

> 5

14. What is a correct Pandas method for returning the last rows?

> tail()

15. What is a correct Pandas method for removing rows that contains empty cells?

> dropna()

16. True or false: by default, the Pandas dropna() method returns a new DataFrame, and will not change the original.

> True

17. What is a correct method to fill empty cells with a new value?

> fillna()

18. What is a correct method to plot (draw) diagrams from the data in a DataFrame?

> df.plot()

19. Mean, Median, Mode. Which one returns the value in the middle?

> median()

20. Mean, Median, Mode. Which one returns the value that appears most frequently?

> mode()

22. What is a correct method to remove duplicates from a Pandas DataFrame?

> df.drop_duplicates()

23. What is a correct method to discover if a row is a duplicate?

> df.duplicated()

24. What is a correct method to find relationships between column in a DataFrame?

> df.corr()

21. Mean, Median, Mode. Which one returns the average value?

> mean()

Which of the following library has DataFrame object?

☑ Pandas

Which of the following is the correct way to import a library, eg Pandas?

☑ import pandas as pd

What is the method of DataFrame object to import a csv file?

☑ from_csv()

Which of the following attributes of a Data Frame return a list of column names of this DataFrame?

☑ Columns

Which of the following can slice 'Close' from '2015-01-01' to '2016-12-31' from data, which is a DataFrame object?

☑ data.loc[' 2015-01-01':'2016-12-31 'Close']

What is the method of DataFrame to plot a line chart?

☑ plot()

Suppose you have a DataFrame - data, which contains columns 'Open', 'High', 'Low', 'Close', 'Adj Close' and 'Volume' of Microsoft's stock.

What does data[['Open', 'Low']] return?

☑ Columns 'Open' and 'Low'

Which of the following syntax calculates the Price difference, (ie 'Close' of tomorrow – 'Close' of today)?

☑ ms['Close'].shift(-1 ms['Close'].shift(-1)

9. Suppose you have a DataFrame - ms , which contains the daily data of 'Open', 'High', 'Low', 'Close', 'Adj Close' and 'Volumn' of Microsoft's stock.

What is the method of DataFrame to calculate the 60 days moving average?

☑ rolling(60).mean()

10. **Which of the following idea(s) is/are correct to the simple trading strategy that we introduced in the lecture video?**

☑

- If fast signal is larger than slow signal. this indicates an upward trend at the current moment
- Use longer moving average as slow signal and shorter moving average as fast signal

Random variables and distribution

1. Roll two dice and X is the sum of faces values. If we roll them 5 times and get 2,3,4,5,6

Which of the following is/are true about X?

☑ X is a random variable

Roll two dice and X is the sum of faces values. If we roll them 5 times and get 2,3,4,5,6

X is a _____ random variable.

☑ Discrete

Why do we use relative frequency instead of frequency?

☑ Relative frequency can be used to compare the ratio of values between difference collections with difference number of values

What do we know about X?

☑ We have 5 observations of X

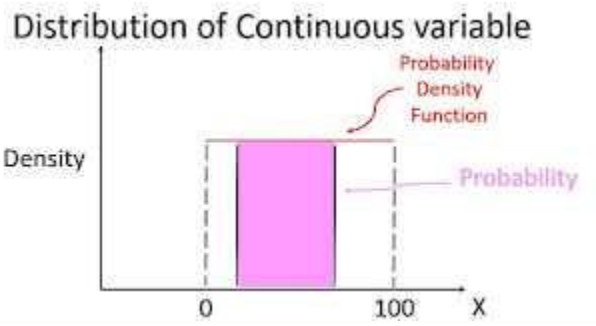
Roll two dice and X is the sum of faces values. If we roll them 5 times and get 2,3,4,5,6. What can we say about relative frequency when we have large number of trials?

☑ Relative frequency becomes approximately the distribution of the corresponding random variable

What is the notion of "95% Value at Risk" ?

☑ 95% VaR measures the amount of investment you can lose, at the worst 5% scenarioa

The calculation of continuous random variable is based on the probability density function.



Given a probability density function, f(x) = 1/100, what is the probability

P(10<X<20), where X~Uniform [0, 100]?

☑ (20- 10) * 1/100

What methods should we use to get the cdf and pdf of normal distribution?

☑ norm.cdf() and norm.pdf() from scipy.stats

Which additional library should we import when we want to calculate log daily return specifically?

☑ Numpy

What is the distribution of stock returns suggested by Fama and French in general?

☑ Close to normal distribution but with fat tail

Sampling and Inference

What is true about sample and population?

☑ Sample is a subset of population which is randomly draw from population

You have a DataFrame called 'data' which has only one column 'population'.
data = pd.DataFrame()
data['population'] = [47, 48, 85, 20, 19, 13, 72, 16, 50, 60]
How to draw sample with sample size =5, from a 'population' with replacement?

☑ data['population'].sample(5, replace=True)

Why is the degrees of freedom n-1 in sample variance?

☑ The degrees of freedom in sample variance is constrained by the sample mean

What does Central Limit Theorem tell you about the distribution of sample mean?

☑ The distribution of sample mean follows normal distribution with very large sample size follows normal distribution regardless of the population distribution

Suppose we have 3 independent normal random variables X1, X2 and X3:
What is the distribution of X1 + X2 + X3?

☑ Mean and variance of X1, X2 and X3 are added up

$N(3\mu, 3\sigma^2)$

Why do we need to standardize sample mean when making inference?

☑ The standardized distribution of sample mean follows N(0,1) which is easier to make inference

What can a 95% confidence interval of daily return of an investment tell you?

☑ With 95% chance this interval will cover the mean of daily return

When do you reject a null hypothesis with alternative hypothesis $\mu>0$ with significance level α?

- p value is smaller than α
- z < z_(1-α)

When doing analysis of stock return, you notice that with 95% confidence interval, the upper bound and lower bound are negative.

Base on this data, what can you tell about this stock?

☑ There is 95% chance of which the mean return of this stock is negative

Linear Regression Models for Financial Analysis

Why do you use coefficient of correlation, instead of covariance, when calculating the association between two random variables ?

☑ Covariance can be affected by the variance of individual variables, but coefficient of correlation is rescaled by variance of both variables

What is the range and interpretation of coefficient of correlation?

☑ From -1 to 1, -1 means perfect negative linear relationship and 1 means perfect positive linear relationship

How to check if a linear regression model violates the independence assumption?

☑ Durbin Watson test

If any of the assumptions of linear regression model are violated, we cannot use this model to make prediction.

☑ False

What does it mean if you have a strategy with maximum drawdown of 3%?

☑ During the trading period, the maximum **drop** from the previous peak of your portfolio value is 3%

How can you check the consistency of your trading strategy?

☑ Define some metric for evaluating your strategy, eg Sharpe Ratio, maximum drawdown. Then split your data into train set and test set and check if your strategy can generate positive return using both train set and **test set**

Matplotlib Plotting with pyplot

1. **Which of the following does not visualize data?**
   a. Charts
   b. Maps
   c. Shapes (correct)
   d. Graphs
1. **Which of the following type of chart is not supported by pyplot?**
   a. Histogram
   b. Boxplot
   c. Pie
   d. All are correct (correct)
1. **In the given chart, box surrounded with red border is called**

a. Data series
b. Chart Title
c. Markers
d. Legend (correct)

To display histogram with well-defined edge we can write

a. df.plot(type = 'hist', edge = 'red')
b. df.plot(type = 'hist', edgecolor = 'red') (correct)
c. df.plot(type = 'hist', line = 'red')
d. df.plot(type = 'hist', linecolor = 'red')

plot which is used to given statistical summary is

a. Bar
b. Line
c. Histogram
d. Box plot (correct)

Which of the following is not the parameter of pyplot's plot() method?

a. Marker
b. Lineheight (correct)
c. Linestyle
d. Color

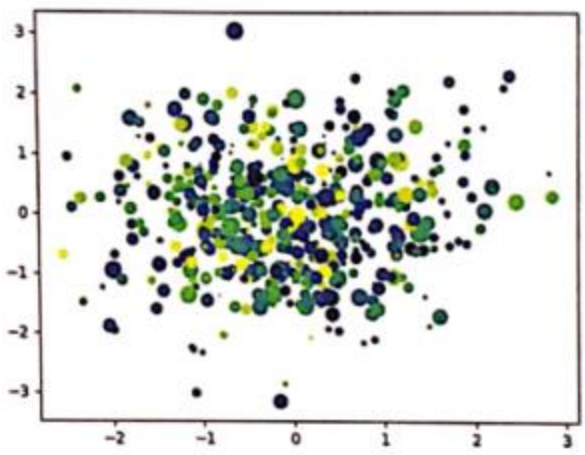
To compare data we can use ____ chart

a. Line
b. Bar (correct)
c. Pie
d. Scatter

Which of the following chart element is used to identify data series by its color patterns?

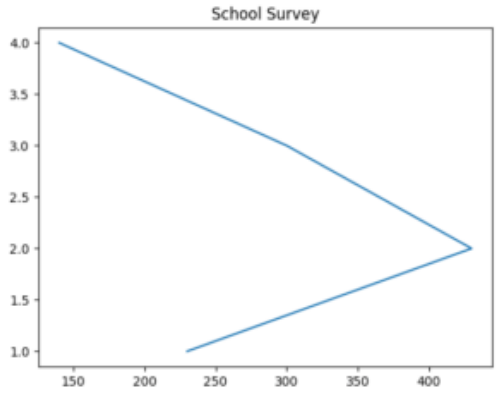
a. Chart title
b. Legend (correct)
c. Marker
d. Data labels

Identify the following chart's type



a. Scatter Chart (correct)
b. Bubble Chart
c. Pie Chart
d. Frequency Polygon

Look at the following graph and select appropriate code to obtain this output. (Please assume that pandas and matplotlib is already imported)



d. zone=[1,2,3,4]
schools = [230,430,300,140]
plt.plot(schools,zone)
plt.title("School Survey")
plt.show()

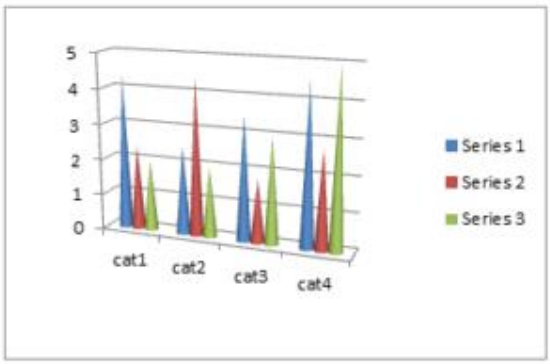
Which of the following is incorrect regarding Data Visualization?
a. Data visualization can be done using Matplotlib library in python.
b. Visualizing large and complex data does not produce effective results. (correct)
c. Data visualization is immensely useful in data analysis.
d. Decision makers use data visualization to understand business problems easily and build strategies.

Matplotlib is _____ plotting library
> 2D

Data _____ refers to graphical representation of data.
> Visualization

1.consider the code given below:
plt.bar(cities, population, color = ['r','g','b','m'])
what will be color of last bar?
a. Magenta (correct)
b. Green
c. Blue

Following chart is type of



> Bar chart

Raj has written following code to create a bar chart-

```
Import matplotlib.pyplot as plt
weekdays  = ['sun','mon','tue']
sale = [1234,3312,2541]
plt.bar(weekdays, sale)
```

But when executing this code he is not able to display any chart. What would you suggest him to execute code and display chart successfully?

a. Instead of using bar() he should use plot() method
b. He should add show() function at the end of code (correct)
c. He should use numpy array to store weekdays and sale data

The interface of Matplotlib used for data visualization is

a. Seaborn
b. Anaconda
c. matlab
d. pyplot (correct)

the best suitable real life example of data visualization is:
a. Percent of population of by age group in India
b. Google Analytics
d. Both a and b

Which of the following is correct code to plot line chart with dotted linestyle?

a. Plt.plot(x,y)
Plt.linestyle = “dotted”
Plt.show()

b. Plt.plot(x,y)
Plt.linestyle(“dotted”)
Plt.show()

c. Plt.plot(x,y, linestyle = “dotted”) (correct)
Plt.show()

1. **in the given chart, box surrounded with red border is called**

b. X label

Which of the following is correct syntax to create histogram with bins specified?

a. Plt.hist(x, bins=10)
b. Plt.hist(x,bins = [10,11,12,13,14])
c. Plt.hist(x, bins = range(10,15))
d. All are correct(correct)

If you install python IDLE, matplotlib is installed automatically.
b. False

The command to install Matplotlib library in python is
d. Pip install matplotlib

in the given chart, box surrounded with red border is called

c. Y label

Select the correct statement to display horizontal box plot
c. Plt.boxplot(data, vert = False)(correct)
b. Plt.boxplot(data, horiz = True)

Which of the following thing can be data in Pandas?

a) a python dict
b) an ndarray
c) a scalar value
d) all of the mentioned

Answer: d
Explanation: The passed index is a list of axis labels.

2. Point out the correct statement.
a) If data is a list, if index is passed the values in data corresponding to the labels in the index will be pulled out
b) NaN is the standard missing data marker used in pandas
c) Series acts very similarly to a array
d) None of the mentioned
Answer: b
Explanation: If data is a dict, if index is passed the values in data corresponding to the labels in the index will be pulled out.

3. The result of an operation between unaligned Series will have the _____ of the indexes involved.
a) intersection
b) union
c) total
d) all of the mentioned
Answer: b
Explanation: If a label is not found in one Series or the other, the result will be marked as missing NaN.

4. Which of the following input can be accepted by DataFrame?
a) Structured ndarray
b) Series

c) DataFrame
d) All of the mentioned
Answer: d
Explanation: DataFrame is a 2-dimensional labeled data structure with columns of potentially different types.

5. Point out the wrong statement.
a) A DataFrame is like a fixed-size dict in that you can get and set values by index label
b) Series can be be passed into most NumPy methods expecting an ndarray
c) A key difference between Series and ndarray is that operations between Series automatically align the data based on label
d) None of the mentioned

Answer: a
Explanation: A Series is like a fixed-size dict in that you can get and set values by index label.

6. Which of the following takes a dict of dicts or a dict of array-like sequences and returns a DataFrame?
a) DataFrame.from_items
b) DataFrame.from_records
c) DataFrame.from_dict
d) All of the mentioned

Answer: a
Explanation: DataFrame.from_dict operates like the DataFrame constructor except for the orient parameter which is 'columns' by default.

7. Series is a one-dimensional labeled array capable of holding any data type.
a) True
b) False

Answer: a
Explanation: The axis labels are collectively referred to as the index.

8. Which of the following works analogously to the form of the dict constructor?
a) DataFrame.from_items
b) DataFrame.from_records
c) DataFrame.from_dict
d) All of the mentioned

Answer: a
Explanation: DataFrame.from_records takes a list of tuples or an ndarray with structured dtype.

9. Which of the following operation works with the same syntax as the analogous dict operations?
a) Getting columns
b) Setting columns
c) Deleting columns
d) All of the mentioned

Answer: d
Explanation: You can treat a DataFrame semantically like a dict of like-indexed Series objects.

10. If data is an ndarray, index must be the same length as data.
a) True
b) False

Answer: a
Explanation: If no index is passed, one will be created having values [0, …, len(data) – 1].

Pandas

1. All pandas data structures are ___ mutable but not always _____mutable.
a) size, value
b) semantic, size
c) value, size
d) none of the mentioned
Answer: c
Explanation: The length of a Series cannot be changed.
2. Point out the correct statement.
a) Pandas consist of set of labeled array data structures
b) Pandas consist of an integrated group by engine for aggregating and transforming data sets
c) Pandas consist of moving window statistics
d) All of the mentioned
Answer: d
Explanation: Some elements may be close to one another according to one distance and farther away according to another.
3. Which of the following statement will import pandas?
a) import pandas as pd
b) import panda as py
c) import pandaspy as pd
d) all of the mentioned
Answer: a
Explanation: You can read data from a CSV file using the read_csv function.
4. Which of the following object you get after reading CSV file?
a) DataFrame
b) Character Vector
c) Panel
d) All of the mentioned
Answer: a
Explanation: You get columns out of a DataFrame the same way you get elements out of a dictionary.

5. Point out the wrong statement.
a) Series is 1D labeled homogeneously-typed array
b) DataFrame is general 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed columns
c) Panel is generally 2D labeled, also size-mutable array
d) None of the mentioned

Answer: c
Explanation: Panel is generally 3D labeled.

6. Which of the following library is similar to Pandas?
a) NumPy
b) RPy
c) OutPy
d) None of the mentioned
Answer: a
Explanation: NumPy is the fundamental package for scientific computing with Python.

7. Panel is a container for Series, and DataFrame is a container for dataFrame objects.
a) True
b) False
Answer: b
Explanation: DataFrame is a container for Series, and panel is a container for dataFrame objects.

8. Which of the following is prominent python “statistics and econometrics library”?
a) Bokeh
b) Seaborn
c) Statsmodels
d) None of the mentioned
Answer: c
Explanation: Bokeh is a Python interactive visualization library for large datasets that natively uses the latest web technologies.

9. Which of the following is a foundational exploratory visualization package for the R language in pandas ecosystem?
a) yhat
b) Seaborn
c) Vincent
d) None of the mentioned
Answer: a
Explanation: It has great support for pandas data objects.

10. Pandas consist of static and moving window linear and panel regression.
a) True
b) False
Answer: a
Explanation: Time series and cross-sectional data are special cases of panel data.

1. The plot method on Series and DataFrame is just a simple wrapper around _____
a) gplt.plot()
b) plt.plot()
c) plt.plotgraph()
d) none of the mentioned
Answer: b
Explanation: If the index consists of dates, it calls gcf().autofmt_xdate() to try to format the x-axis nicely.

2. Point out the correct combination with regards to kind keyword for graph plotting.
a) ‘hist’ for histogram
b) ‘box’ for boxplot
c) ‘area’ for area plots
d) all of the mentioned
Answer: d
Explanation: The kind keyword argument of plot() accepts a handful of values for plots other than the default Line plot.

3. Which of the following value is provided by kind keyword for barplot?
a) bar
b) kde
c) hexbin
d) none of the mentioned
Answer: a
Explanation: bar can also be used for barplot.

4. You can create a scatter plot matrix using the _____ method in pandas.tools.plotting.
a) sca_matrix
b) scatter_matrix
c) DataFrame.plot
d) all of the mentioned

Answer: b
Explanation: You can create density plots using the Series/DataFrame.plot.

5. Point out the wrong combination with regards to kind keyword for graph plotting.
a) ‘scatter’ for scatter plots
b) ‘kde’ for hexagonal bin plots
c) ‘pie’ for pie plots
d) none of the mentioned

Answer: b
Explanation: kde is used for density plots.

6. Which of the following plots are used to check if a data set or time series is random?
a) Lag
b) Random
c) Lead
d) None of the mentioned

Answer: a
Explanation: Random data should not exhibit any structure in the lag plot.

7. Plots may also be adorned with error bars or tables.
a) True
b) False

Answer: a
Explanation: There are several plotting functions in pandas.tools.plotting.

8. Which of the following plots are often used for checking randomness in time series?
a) Autocausation
b) Autorank
c) Autocorrelation
d) None of the mentioned

Answer: c
Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

9. _____ plots are used to visually assess the uncertainty of a statistic.
a) Lag
b) RadViz
c) Bootstrap
d) None of the mentioned

Answer: c
Explanation: Resulting plots and histograms are what constitutes the bootstrap plot.

Andrews curves allow one to plot multivariate data.
a) True
b) False
Answer: a
Explanation: Curves belonging to samples of the same class will usually be closer together and form larger structures

Python MCQ (Multi Choice Questions)

Q. What is the maximum possible length of an identifier?
a. 16
b. 32
c. 64
d. None of these above

Answer: (d) None of these above
Explanation: The maximum possible length of an identifier is not defined in the python language. It can be of any numbe

Q. Who developed the Python language?
a. Zim Den
b. Guido van Rossum
c. Niene Stom
d. Wick van Rossum

Answer: (b) Guido van Rossum
Explanation: Python language was developed by Guido van Rossum in the Netherlands.

Q. In which year was the Python language developed?

a. 1995
b. 1972
c. 1981
d. 1989

Answer: (d) 1989
Explanation: Python language was developed by Guido van Rossum in 1989.

Q. In which language is Python written?
a. English
b. PHP
c. C
d. All of the above

Answer: (b) C
Explanation: Python is written in C programming language, and it is also called CPython.

Q. Which one of the following is the correct extension of the Python file?
a. .py
b. .python
c. .p
d. None of these

Answer: (a) .py
Explanation: ".py" is the correct extension of the Python file.

Q. In which year was the Python 3.0 version developed?

a. 2008
b. 2000
c. 2010
d. 2005

Q. What do we use to define a block of code in Python language?
a. Key
b. Brackets
c. Indentation
d. None of these

Answer: (c) Indentation
Explanation: Python uses indentation to define blocks of code. Indentations are simply spaces or tabs used as an indicat

Q. Which character is used in Python to make a single line comment?
a. /
b. //
c. #
d. !

Answer: (c) #
Explanation: "#" character is used in Python to make a single-line comment.

Q. Which of the following statements is correct regarding the object-oriented programming concept in Python?
a. Classes are real-world entities while objects are not real
b. Objects are real-world entities while classes are not real
c. Both objects and classes are real-world entities
d. All of the above

Answer: (b) Objects are real-world entities while classes are not real
Explanation: None

Q. Which of the following statements is correct in this python code?

```
1. class Name:
2. def __init__(javatpoint):
3. javajavatpoint = java
4. name1=Name("ABC")
5. name2=name1
```

a. It will throw the error as multiple references to the same object is not possible
b. id(name1) and id(name2) will have same value
c. Both name1 and name2 will have reference to two different objects of class Name
d. All of the above

Answer: (b) id(name1) and id(name2) will have same value
Explanation: "name1" and "name2" refer to the same object, so id(name1) and id(name2) will have the same value.

Q. What is the method inside the class in python language?
a. Object
b. Function
c. Attribute
d. Argument

Answer: (b) Function
Explanation: Function is also known as the method.

Q. Which of the following declarations is incorrect?
a. _x = 2
b. __x = 3
c. __xyz__= 5
d. None of these

Answer: (d) None of these
Explanation: All declarations will execute successfully but at the expense of low readability.

Q. Why does the name of local variables start with an underscore discouraged?
a. To identify the variable
b. It confuses the interpreter
c. It indicates a private variable of a class
d. None of these

Answer: (c) It indicates a private variable of a class
Explanation: Since there is no concept of private variables in Python language, the major underscore is used to denote va

Q. Which of the following is not a keyword in Python language?
a. val
b. raise
c. try
d. with

Answer: (a) val
Explanation: "val" is not a keyword in python language.

Q. Which of the following statements is correct for variable names in Python language?
a. All variable names must begin with an underscore.
b. Unlimited length
c. The variable name length is a maximum of 2.
d. All of the above

Answer: (b) Unlimited length
Explanation: None

Q. Which of the following declarations is incorrect in python language? a. xyzp = 5,000,000
b. x y z p = 5000 6000 7000 8000
c. x,y,z,p = 5000, 6000, 7000, 8000
d. x_y_z_p = 5,000,000

Answer: (b) x y z p = 5000 6000 7000 8000

Explanation: Spaces are not allowed in variable names.

Answer: (a) 2008
Explanation: Python 3.0 version was developed on December 3, 2008.

Data Science

1. Artifical intelligence comprises of
A. Natural language processing, computer vision but not robtics and machine learning
B. Natural language processing and robotics but not computer vision and machine learning
C. Natural language processing, computer vision, robotics and machine learning
D. NA

2 Deep Learning is a sub area under machine learning?
A. TRUE
B. FALSE

3. Which of the following is not true about unsupervised learning:
A. It finds clusters of the data
B. Finding interesting co-relation and coordinates with the data
C. Finding the annotate strings and predicting time series
D. Finding low dimensional representations of data

4. If a row of a data matrix represents an object, a column usually represents?
A. Attribute
B. Record
C. Table name
D. Entity

5. College name of a student can be considered as an attribute of type?
A. Nominal
B. Ordinal
C. Interval
D. Ratio

6. Which of the following operations can be performed on nominal attributes?
A. Distinctness
B. Order
C. Both of the above
D. none of the above

7. Which of the following is an example of discrete attribute?
A. Thickness of a book in centimetre

B. Weight of a book in grams

C. number of words in a book

D. none of the above

8. Discretization of attributes refers to:

A. Deleting data objects

B. Deleting data attributes

C. Modifying nature of attribute values

D. Removing noisy data objects

9. Which of the following can be considered as a Data Science Project?

A. Registering for an online course

B. Online money payment through a bank

C. Predicting if a student will pass an online course

D. Downloading the certificate from course website

10. Which of the below is a proper sequence of steps in data science projects? 1) Creating a target data set: data selection 2) Visualiz

A. a,b,c,d,e,f,g

B. a,c,b,f,d,g,e

C. a,c,b,d,f,g,e

11. while visualizing a distribution, what does a large standard deviation suggest?

A. Data and values are widely distributed and that the mean may not be a reliable measure of central tendency.

B. The values are not widely distributed and the median would be an unreliable measure of the central tendency.

C. Values are not normally distributed.

D. All of the measures of central tendency would be reliable.

12. while visualizing a distribution, what does a large standard deviation suggest?

A. Data and values are widely distributed and that the mean may not be a reliable measure of central tendency.

B. The values are not widely distributed and the median would be an unreliable measure of the central tendency.

C. Values are not normally distributed.

D. All of the measures of central tendency would be reliable.

13. Consider the following statements and answer the question: Statement I: A histogram represents the frequencies of all x values v

A. Statement I is true and Statement II is false

B. Statement II is true and Statement I is false

C. Both the statements are true

D. Both the statements are false

14. Histogram and Frequency Polygon represent the same things Histogram and Frequency Polygon represent the same things

A. TRUE

B. FALSE

15. Which of the following approaches can be used for dimension reduction? I. Domain knowledge II. Data exploration techniques III

A. Only II, IV and V

B. Only I,III and IV

C. Only I,IV and V

D. I, II, III, IV and V

16. If the Independant and depandant variables are continuous then which test can be used for checking the possibility of dimesionli

A. Correlation

B. t-test

C. ChiSquared test

D. Anova

17. Conversion of Continuous to Categorical data using techniques like binning is a technique used in

A. Feature Enginering

B. Visualization

C. Both

18. Dimensionality Reduction is done I) When dimensionality increases, data becomes decreasingly sparse in the space that it occupi

A. I is true

B. II is true

C. Both are true

D. None are true

19. Which of the following are examples of Descriptive statistics

A. Mean,Median

B. Range,Statdard Devaiation

C. Frequency, percentage

D. All the above

20. Variance and Standard deviation are both measures of variability

A. TRUE

B. FALSE

21. Z-Score shows:

A. Number of Std deviations, an data in a sample is away from its sample mean

B. Number of data in a sample having values, above sample data for whicn z-score is calculated

C. Number of data in a sample having values, below sample data for whicn z-score is calculated

22. Which metric qualifies the shape of a distribution

A. Variation

B. Kurtosis

C. Skewness

23. In right Skewed distributions Mean is less than median

A. TRUE

B. FALSE

24. Which metric measures, how sharp the peak of a distribution is.

A. Variation

B. Kurtosis

C. Skewness

25. Eucliendean distance doesn't follow pythogorous theorm.

A. TRUE

B. FALSE

26. Which one of the following is least sensitive to outliers?

A. Mean

B. Mode

C. Median

D. None of the above

27. In any probablity distribution graph, Y-axis is always frequency.

A. TRUE

B. FALSE

28. Which one of the following is the benefit of using simple random sampling?

A. Informants won't refuse to participate

B. Interviewers can choose respondent freely

C. We can calculate the accuracy of the results

D. The results are always representative

29. The value of the t statistic for the hypotheses test is (final value is rounded off to 2 decimal places)

A. 3.59

B. 69.73

C. 67.97

30. What does a dummy variable regression analysis examine?

A. Relationship between one continuous dependent and one continuous independent variable

B. Relationship between one categorical dependent and one continuous independent variable

C. Relationship between one continuous dependent and one categorical independent variable

D. Relationship between one continuous dependent and one classified variable

31. What assumptions are made when we use the t-distribution to perform a hypothesis test?

A. The underline population has a constant variance

B. The underline population has a non-symmetrical distribution

C. The underline population follows a normal distribution

D. All of the above

32. Lottery ticket is an example of sampling with replacement

A. TRUE

B. FALSE

33. Sampling errors are caused by

A. Actual Process of sampling

B. Factors not related to the sampling process

C. All of the above

34. The discrepancy between a population statistic and its population parameter is called sampling error.

A. TRUE

B. FALSE

35. Coefficient of correlation closer to -1 indicated weak assosiation between the two variables.
A. TRUE

B. FALSE

36. Pick the wrong statement:
A. Partitioning involves removing biased from the data.

B. Evaluation of the test set results in overfitting when the same mode is used.

C. Partitioning creates multiple subsets of data used for data visualisation.

D. Partition is used for validation to find tune the data and improve the model.

37. If the model shows an overfit what should be done to reduce the over fit:

A. Increase the Train split %

B. Boot Strap sample the train data

C. Both are true

D. None of the above

38. "Stratified partitioning of trains and test samples doent not improve the model performance "
A. TRUE

B. FALSE

39. The coefficient of determination (R2)value is always bounded between
A. [0,1]

B. [1,2]

C. [-1,+1]

D. [-2,+2]

40. Assuming 95 % confidence level, the confidence interval for the estimation of intercept for the linear model built using the milk p
A. [0.45, 5.54]

B. [0.70,0.91]

C. [7.93, 21.05]

D. [4.51, -6.50]

41. The number of outlier(s) from the residual plot bounded between 2σ limits
A. 1

B. 10

C. 11

D. None

42. "For the same milk production dataset, the observed F statistic of the full linear regression model is 245.1 . The theoretical F statistic at 0.1%, 1% and 5% significance level are 15.38, 8.29 and 4.41 respectively. The observed F statistic will be accepted at "
A. 5 % significance level

B. 1 % significance level

C. 0.1 % significance level

D. All of the above significance levels

43. For a univariate linear regression Standardized residuals have
A. T distribution with n-2 df

B. Normal distribution with n-2 df

C. T distribution with n-1 df

D. Normal distribution with n-1 df

44. "Which of the following correlation coefficient is used to measure the association between two ordinal variables? "

A. Pearson and Spearman

B. Spearman

C. Kendall

D. Spearman and Kendall

45. In the equation Minutes=4.16+15.51Units :
A. Minutes is the independent variable and Units is the dependent variable

B. Units is the independent variable and Minutes is the dependent variable

C. Both Units and Minutes are the independent variables

D. Both Units and Minutes are the dependent variables

46. In data science projects, the prediction error for performance evaluation is typically computed on:
A. Training partition

B. Validation partition

C. Test partition

D. None of the above

47. Which of the following predictive accuracy metrics indicate percentage deviation from actual values?
A. MAE

B. RMSE

C. MAPE

D. SSE

48. The primary objective of predictive modeling in multiple linear regression is to:
A. Fit the data closely

B. Estimate values of outcome variable for new records accurately

C. Understand the relationship between outcome variable and predictors

D. All of the above

NLP Questions

..

Q1. Which of the following techniques can be used for keyword normalization in NLP, the process of converting a keyword into its base form?

a. Lemmatization

b. Soundex

c. Cosine Similarity
d. N-grams
Answer : a) Lemmatization helps to get to the base form of a word, e.g. are playing -> play, eating -> eat, etc.Other options are meant for different purposes.

Q2. Which of the following techniques can be used to compute the distance between two word vectors in NLP?

a. Lemmatization
b. Euclidean distance
c. Cosine Similarity
d. N-grams
Answer : b) and c)
Distance between two word vectors can be computed using Cosine similarity and Euclidean Distance. Cosine Similarity establishes a cosine angle between the vector of two words. A cosine angle close to each other between two word vectors indicates the words are similar and vice a versa.
E.g. cosine angle between two words "Football" and "Cricket" will be closer to 1 as compared to angle between the words "Football" and "New Delhi"

Q3. What are the possible features of a text corpus in NLP?

a. Count of the word in a document
b. Vector notation of the word
c. Part of Speech Tag
d. Basic Dependency Grammar
e. All of the above
Answer : e)All of the above can be used as features of the text corpus.

Q4. You created a document term matrix on the input data of 20K documents for a Machine learning model. Which of the following can be used to reduce the dimensions of data?

..
1. Keyword Normalization
2. Latent Semantic Indexing
3. Latent Dirichlet Allocation
a. only 1
b. 2, 3
c. 1, 3
d. 1, 2, 3
Answer : d)

Q5. Which of the text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection, and object detection in NLP.

a. Part of speech tagging
b. Skip Gram and N-Gram extraction
c. Continuous Bag of Words
d. Dependency Parsing and Constituency Parsing
Answer : d)

Q6. Dissimilarity between words expressed using cosine similarity will have values significantly higher than 0.5

a. True
b. False
Answer : a)

Q7. Which one of the following are keyword Normalization techniques in NLP

a. Stemming
b. Part of Speech
c. Named entity recognition
d. Lemmatization

Answer : a) and d)
Part of Speech (POS) and Named Entity Recognition(NER) are not keyword Normalization techniques. Named Entity help you extract Organization, Time, Date, City, etc..type of entities from the given sentence, whereas Part of Speech helps you extract Noun, Verb, Pronoun, adjective, etc..from the given sentence tokens.

Q8. Which of the below are NLP use cases?

a. Detecting objects from an image
b. Facial Recognition
c. Speech Biometric
d. Text Summarization

..

Answer : (d)
a) And b) are Computer Vision use cases, and c) is Speech use case.
Only d) Text Summarization is an NLP use case.

Q9. In a corpus of N documents, one randomly chosen document contains a total of T terms and the term "hello" appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-documentfrequency), if the term "hello" appears in approximately one-third of the total documents?

a. KT * Log(3)
b. T * Log(3) / K
c. K * Log(3) / T
d. Log(3) / KT

Answer : (c)
formula for TF is K/T
formula for IDF is log(total docs / no of docs containing "data")
= log(1 / (⅓))
= log (3)
Hence correct choice is Klog(3)/T

Q10. In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

a. Term Frequency (TF)
b. Inverse Document Frequency (IDF)
c. Word2Vec
d. Latent Dirichlet Allocation (LDA)

Answer : b)

Q11. In NLP, The process of removing words like "and", "is", "a", "an", "the" from a sentence is called as

a. Stemming
b. Lemmatization
c. Stop word
d. All of the above

Answer : c) In Lemmatization, all the stop words such as a, an, the, etc.. are removed. One can also define custom stop words for removal.

..

Q12. In NLP, The process of converting a sentence or paragraph into tokens is referred to as Stemming

a. True
b. False

Answer : b) The statement describes the process of tokenization and not stemming, hence it is False.

Q13. In NLP, Tokens are converted into numbers before giving to any Neural Network

a. True
b. False
Answer : a) In NLP, all words are converted into a number before feeding to a Neural Network.
Q14 Identify the odd one out
a. nltk
b. scikit learn
c. SpaCy
d. BERT
Answer : d) All the ones mentioned are NLP libraries except BERT, which is a word embedding
Q15 TF-IDF helps you to establish?
a. most frequently occurring word in the document
b. most important word in the document
Answer : b) TF-IDF helps to establish how important a particular word is in the context of the document corpus. TF-IDF takes into account the number of times the word appears in the document and offset by the number of documents that appear in the corpus.
● TF is the frequency of term divided by a total number of terms in the document.
● IDF is obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient.
..
● Tf.idf is then the multiplication of two values TF and IDF.
Q16 In NLP, The process of identifying people, an organization from a given sentence, paragraph is called
a. Stemming
b. Lemmatization
c. Stop word removal
d. Named entity recognition
Answer : d)
Q17 Which one of the following is not a pre-processing technique in NLP
a. Stemming and Lemmatization
b. converting to lowercase
c. removing punctuations
d. removal of stop words
e. Sentiment analysis
Answer : e) Sentiment Analysis is not a pre-processing technique. It is done after pre-processing and is an NLP use case. All other listed ones are used as part of statement pre-processing.
Q18 In text mining, converting text into tokens and then converting them into an integer or floating-point vectors can be done using
a. CountVectorizer
b. TF-IDF
c. Bag of Words
d. NERs
Answer : a) CountVectorizer helps do the above, while others are not applicable.
text =[“Rahul is an avid writer, he enjoys studying understanding and presenting. He loves to play”]
vectorizer = CountVectorizer()
vectorizer.fit(text)
vector = vectorizer.transform(text)
..
print(vector.toarray())
output
[[1 1 1 1 2 1 1 1 1 1 1 1 1]]
The second section of the interview questions covers advanced NLP techniques such as Word2Vec, GloVe word embeddings, and advanced models such as GPT, ELMo, BERT, XLNET

based questions, and explanations.

Q19. In NLP, Words represented as vectors are called as Neural Word Embeddings

a. True

b. False

Answer : a) Word2Vec, GloVe based models build word embedding vectors that are multidimensional.

Q20. In NLP, Context modeling is supported with which one of the following word embeddings

1. a. Word2Vec
2. b) GloVe
3. c) BERT
4. d) All of the above

Answer : c) Only BERT (Bidirectional Encoder Representations from Transformer) supports context modelling where the previous and next sentence context is taken into consideration. In Word2Vec, GloVe only word embeddings are considered and previous and next sentence context is not considered.

Q21. In NLP, Bidirectional context is supported by which of the following embedding

a. Word2Vec

b. BERT

c. GloVe

d. All the above

Answer : b) Only BERT provides a bidirectional context. The BERT model uses the previous and the next sentence to arrive at the context.Word2Vec and GloVe are word embeddings, they do not provide any context.

Q22. Which one of the following Word embeddings can be custom trained for a specific subject in NLP

a. Word2Vec

b. BERT

..

c. GloVe

d. All the above

Answer : b) BERT allows Transform Learning on the existing pre-trained models and hence can be custom trained for the given specific subject, unlike Word2Vec and GloVe where existing word embeddings can be used, no transfer learning on text is possible.

Q23. Word embeddings capture multiple dimensions of data and are represented as vectors

a. True

b. False

Answer : a)

Q24. In NLP, Word embedding vectors help establish distance between two tokens

a. True

b. False

Answer : a) One can use Cosine similarity to establish distance between two vectors represented through Word Embeddings

Q25. Language Biases are introduced due to historical data used during training of word embeddings, which one amongst the below is not an example of bias

a. New Delhi is to India, Beijing is to China

b. Man is to Computer, Woman is to Homemaker

Answer : a)

Statement b) is a bias as it buckets Woman into Homemaker, whereas statement a) is not a biased statement.

Q26. Which of the following will be a better choice to address NLP use cases such

as semantic similarity, reading comprehension, and common sense reasoning
a. ELMo
b. Open AI's GPT
c. ULMFit
Answer : b) Open AI's GPT is able to learn complex pattern in data by using the Transformer models Attention mechanism and hence is more suited for complex use cases such as semantic similarity, reading comprehensions, and common sense reasoning.
Q27. Transformer architecture was first introduced with?
a. GloVe
b. BERT
..
c. Open AI's GPT
d. ULMFit
Answer : c) ULMFit has an LSTM based Language modeling architecture. This got replaced into Transformer architecture with Open AI's GPT
Q28. Which of the following architecture can be trained faster and needs less amount of training data
a. LSTM based Language Modelling
b. Transformer architecture
Answer : b) Transformer architectures were supported from GPT onwards and were faster to train and needed less amount of data for training too.
Q29. Same word can have multiple word embeddings possible with _____?
a. GloVe
b. Word2Vec
c. ELMo
d. nltk
Answer : c) EMLo word embeddings supports same word with multiple embeddings, this helps in using the same word in a different context and thus captures the context than just meaning of the word unlike in GloVe and Word2Vec. Nltk is not a word embedding.
Q30 For a given token, its input representation is the sum of embedding from the token, segment and position embedding
a. ELMo
b. GPT
c. BERT
d. ULMFit
Answer : c) BERT uses token, segment and position embedding.
Q31. Trains two independent LSTM language model left to right and right to left and shallowly concatenates them
a. GPT
b. BERT
c. ULMFit
d. ELMo
..
Answer : d) ELMo tries to train two independent LSTM language models (left to right and right to left) and concatenates the results to produce word embedding.
Q32. Uses unidirectional language model for producing word embedding
a. BERT
b. GPT
c. ELMo
d. Word2Vec
Answer : b) GPT is a unidirectional model and word embedding are produced by training on information flow from left to right. ELMo is bidirectional but shallow. Word2Vec provides simple word embedding.

Q33. In this architecture, the relationship between all words in a sentence is modelled irrespective of their position. Which architecture is this?

a. OpenAI GPT
b. ELMo
c. BERT
d. ULMFit

Answer : c)BERT Transformer architecture models the relationship between each word and all other words in the sentence to generate attention scores. These attention scores are later used as weights for a weighted average of all words' representations which is fed into a fully-connected network to generate a new representation.

Q34. List 10 use cases to be solved using NLP techniques?

● Sentiment Analysis
● Language Translation (English to German, Chinese to English, etc..)
● Document Summarization
● Question Answering
● Sentence Completion
● Attribute extraction (Key information extraction from the documents)
● Chatbot interactions
● Topic classification
● Intent extraction
● Grammar or Sentence correction
● Image captioning
● Document Ranking
● Natural Language inference

Q35. Transformer model pays attention to the most important word in Sentence

a. True
b. False

..

Answer : a) Attention mechanisms in the Transformer model are used to model the relationship between all words and also provide weights to the most important word.

Q36. Which NLP model gives the best accuracy amongst the following?

a. BERT
b. XLNET
c. GPT-2
d. ELMo

Answer : b) XLNET has given best accuracy amongst all the models. It has outperformed BERT on 20 tasks and achieves state of art results on 18 tasks including sentiment analysis, question answering, natural language inference, etc.

Q37. Permutation Language models is a feature of

a. BERT
b. EMMo
c. GPT
d. XLNET

Answer : d) XLNET provides permutation-based language modelling and is a key difference from BERT. In permutation language modeling, tokens are predicted in a random manner and not sequential. The order of prediction is not necessarily left to right and can be right to left. The original order of words is not changed but a prediction can be random.

The conceptual difference between BERT and XLNET can be seen from the following diagram.

Q38. Transformer XL uses relative positional embedding

a. True
b. False

a) Instead of embedding having to represent the absolute position of a word, Transformer XL uses an embedding to encode the relative distance between the words. This embedding is used to

compute the attention score between any 2 words that could be separated by n words before or after.

Q39. What is Naive Bayes algorithm, When we can use this algorithm in NLP?

Naive Bayes algorithm is a collection of classifiers which works on the principles of the Bayes' theorem. This series of NLP model forms a family of algorithms that can be used for a wide range of classification tasks including sentiment prediction, filtering of spam, classifying documents and more.

Naive Bayes algorithm converges faster and requires less training data. Compared to other discriminative models like logistic regression, Naive Bayes model it takes lesser time to train. This algorithm is perfect for use while working with multiple classes and text classification where the data is dynamic and changes frequently.

Q40. Explain Dependency Parsing in NLP?

..

Dependency Parsing, also known as Syntactic parsing in NLP is a process of assigning syntactic structure to a sentence and identifying its dependency parses. This process is crucial to understand the correlations between the "head" words in the syntactic structure.

The process of dependency parsing can be a little complex considering how any sentence can have more than one dependency parses. Multiple parse trees are known as ambiguities. Dependency parsing needs to resolve these ambiguities in order to effectively assign a syntactic structure to a sentence.

Dependency parsing can be used in the semantic analysis of a sentence apart from the syntactic structuring.

Q41. What is text Summarization?

Text summarization is the process of shortening a long piece of text with its meaning and effect intact. Text summarization intends to create a summary of any given piece of text and outlines the main points of the document. This technique has improved in recent times and is capable of summarizing volumes of text successfully.

Text summarization has proved to a blessing since machines can summarise large volumes of text in no time which would otherwise be really time-consuming. There are two types of text summarization:

- Extraction-based summarization
- Abstraction-based summarization

Q42. What is NLTK? How is it different from Spacy?

NLTK or Natural Language Toolkit is a series of libraries and programs that are used for symbolic and statistical natural language processing. This toolkit contains some of the most powerful libraries that can work on different ML techniques to break down and understand human language. NLTK is used for Lemmatization, Punctuation, Character count, Tokenization, and Stemming. The difference between NLTK and Spacey are as follows:

- While NLTK has a collection of programs to choose from, Spacey contains only the bestsuited algorithm for a problem in its toolkit
- NLTK supports a wider range of languages compared to Spacey (Spacey supports only 7 languages)
- While Spacey has an object-oriented library, NLTK has a string processing library
- Spacey can support word vectors while NLTK cannot

Q43. What is information extraction?

Information extraction in the context of Natural Language Processing refers to the technique of extracting structured information automatically from unstructured sources to ascribe meaning to it. This can include extracting information regarding attributes of entities, relationship between different entities and more. The various models of information extraction includes:

- Tagger Module
- Relation Extraction Module
- Fact Extraction Module
- Entity Extraction Module

..
● Sentiment Analysis Module
● Network Graph Module
● Document Classification & Language Modeling Module

Q44. What is Bag of Words?

Bag of Words is a commonly used model that depends on word frequencies or occurrences to train a classifier. This model creates an occurrence matrix for documents or sentences irrespective of its grammatical structure or word order.

Q45. What is Pragmatic Ambiguity in NLP?

Pragmatic ambiguity refers to those words which have more than one meaning and their use in any sentence can depend entirely on the context. Pragmatic ambiguity can result in multiple interpretations of the same sentence. More often than not, we come across sentences which have words with multiple meanings, making the sentence open to interpretation. This multiple interpretation causes ambiguity and is known as Pragmatic ambiguity in NLP.

Q46. What is a Masked Language Model?

Masked language models help learners to understand deep representations in downstream tasks by taking an output from the corrupt input. This model is often used to predict the words to be used in a sentence.

Q48. What are the best NLP Tools?

Some of the best NLP tools from open sources are:
● SpaCy
● TextBlob
● Textacy
● Natural language Toolkit
● Retext
● NLP.js
● Stanford NLP
● CogcompNLP

Q49. What is POS tagging?

Parts of speech tagging better known as POS tagging refers to the process of identifying specific words in a document and group them as part of speech, based on its context. POS tagging is also known as grammatical tagging since it involves understanding grammatical structures and identifying the respective component.

POS tagging is a complicated process since the same word can be different parts of speech depending on the context. The same generic process used for word mapping is quite ineffective for POS tagging because of the same reason.

Q50. What is NES?

Name entity recognition is more commonly known as NER is the process of identifying specific entities in a text document which are more informative and have a unique context. These often denote places, people, organisations, and more. Even though it seems like these entities are
..
proper nouns, the NER process is far from identifying just the nouns. In fact, NER involves entity chunking or extraction wherein entities are segmented to categorise them under different predefined classes. This step further helps in extracting information.

Q51 Explain the Masked Language Model?

Masked language modelling is the process in which the output is taken from the corrupted input. This model helps the learners to master the deep representations in downstream tasks. You can predict a word from the other words of the sentence using this model.

Q52 What is pragmatic analysis in NLP?

Pragmatic Analysis: It deals with outside word knowledge, which means knowledge that is external to the documents and/or queries. Pragmatics analysis that focuses on what was described is reinterpreted by what it actually meant, deriving the various aspects of language that require real-world knowledge.

Q53 What is perplexity in NLP?

The word "perplexed" means "puzzled" or "confused", thus Perplexity in general means the inability to tackle something complicated and a problem that is not specified. Therefore, Perplexity in NLP is a way to determine the extent of uncertainty in predicting some text.

In NLP, perplexity is a way of evaluating language models. Perplexity can be high and low; Low perplexity is ethical because the inability to deal with any complicated problem is less while high perplexity is terrible because the failure to deal with a complicated is high.

Q54 What is ngram in NLP?

N-gram in NLP is simply a sequence of n words, and we also conclude the sentences which appeared more frequently, for example, let us consider the progression of these three words:

- New York (2 gram)
- The Golden Compass (3 gram)
- She was there in the hotel (4 gram)

Now from the above sequence, we can easily conclude that sentence (a) appeared more frequently than the other two sentences, and the last sentence(c) is not seen that often. Now if we assign probability in the occurrence of an n-gram, then it will be advantageous. It would help in making next-word predictions and in spelling error corrections.

Q55 Explain differences between AI, Machine Learning and NLP

..

Q56 Why self-attention is awesome?

“In terms of computational complexity, self-attention layers are faster than recurrent layers when the sequence length n is smaller than the representation dimensionality d, which is most often the case with sentence representations used by state-of-the-art models in machine translations, such as word-piece and byte-pair representations.” — from Attention is all you need

Q57 What are stop words?

Stop words are said to be useless data for a search engine. Words such as articles, prepositions, etc. are considered as stop words. There are stop words such as was, were, is, am, the, a, an, how, why, and many more. In Natural Language Processing, we eliminate the stop words to understand and analyze the meaning of a sentence. The removal of stop words is one of the most important tasks for search engines. Engineers design the algorithms of search engines in such a way that they ignore the use of stop words. This helps show the relevant search result for a query.

Q58 What is Latent Semantic Indexing (LSI)?

..

Latent semantic indexing is a mathematical technique used to improve the accuracy of the information retrieval process. The design of LSI algorithms allows machines to detect the hidden (latent) correlation between semantics (words). To enhance information understanding, machines generate various concepts that associate with the words of a sentence.

The technique used for information understanding is called singular value decomposition. It is generally used to handle static and unstructured data. The matrix obtained for singular value decomposition contains rows for words and columns for documents. This method best suits to identify components and group them according to their types.

The main principle behind LSI is that words carry a similar meaning when used in a similar context. Computational LSI models are slow in comparison to other models. However, they are good at contextual awareness that helps improve the analysis and understanding of a text or a document.

Q60 What are Regular Expressions?

A regular expression is used to match and tag words. It consists of a series of characters for matching strings.

Suppose, if A and B are regular expressions, then the following are true for them:

- If {ε} is a regular language, then ε is a regular expression for it.
- If A and B are regular expressions, then A + B is also a regular expression within the language {A, B}.
- If A and B are regular expressions, then the concatenation of A and B (A.B) is a regular expression.

● If A is a regular expression, then A* (A occurring multiple times) is also a regular expression.

Q61 What are unigrams, bigrams, trigrams, and n-grams in NLP?

When we parse a sentence one word at a time, then it is called a unigram. The sentence parsed two words at a time is a bigram.

When the sentence is parsed three words at a time, then it is a trigram. Similarly, n-gram refers to the parsing of n words at a time.

Example: To understand unigrams, bigrams, and trigrams, you can refer to the below diagram:

Q62 What are the steps involved in solving an NLP problem?

Below are the steps involved in solving an NLP problem:

1. Gather the text from the available dataset or by web scraping

..

2. Apply stemming and lemmatization for text cleaning
3. Apply feature engineering techniques
4. Embed using word2vec
5. Train the built model using neural networks or other Machine Learning techniques
6. Evaluate the model's performance
7. Make appropriate changes in the model
8. Deploy the model

Q63. There have some various common elements of natural language processing. Those elements are very important for understanding NLP properly, can you please explain the same in details with an example?

Answer:

There have a lot of components normally using by natural language processing (NLP). Some of the major components are explained below:

● Extraction of Entity: It actually identifying and extracting some critical data from the available information which help to segmentation of provided sentence on identifying each entity. It can help in identifying one human that it's fictional or real, same kind of reality identification for any organization, events or any geographic location etc.

● The analysis in a syntactic way: it mainly helps for maintaining ordering properly of the available words.

Q64 In the case of processing natural language, we normally mentioned one common terminology NLP and binding every language with the same terminology properly. Please explain in details about this NLP terminology with an example?

Answer:

This is the basic NLP Interview Questions asked in an interview. There have some several factors available in case of explaining natural language processing. Some of the key factors are given below:

● Vectors and Weights: Google Word vectors, length of TF-IDF, varieties documents, word vectors, TF-IDF.

● Structure of Text: Named Entities, tagging of part of speech, identifying the head of the sentence.

● Analysis of sentiment: Know about the features of sentiment, entities available for the sentiment, sentiment common dictionary.

● Classification of Text: Learning supervising, set off a train, set of validation in Dev, Set of define test, a feature of the individual text, LDA.

● Reading of Machine Language: Extraction of the possible entity, linking with an individual entity, DBpedia, some libraries like Pikes or FRED.

Q65 Explain briefly about word2vec

Word2Vec embeds words in a lower-dimensional vector space using a shallow neural network. The result is a set of word-vectors where vectors close together in vector space have similar meanings based on context, and word-vectors distant to each other have differing meanings. For example, apple and orange would be close together and apple and gravity would be relatively far.

..
There are two versions of this model based on skip-grams (SG) and continuous-bag-of-words (CBOW).

Q66 What are the metrics used to test an NLP model?

Accuracy, Precision, Recall and F1. Accuracy is the usual ratio of the prediction to the desired output. But going just be accuracy is naive considering the complexities involved.

Q67 What are some ways we can preprocess text input?

Here are several preprocessing steps that are commonly used for NLP tasks:

● case normalization: we can convert all input to the same case (lowercase or uppercase) as a way of reducing our text to a more canonical form

● punctuation/stop word/white space/special characters removal: if we don't think these words or characters are relevant, we can remove them to reduce the feature space

● lemmatizing/stemming: we can also reduce words to their inflectional forms (i.e. walks → walk) to further trim our vocabulary

● generalizing irrelevant information: we can replace all numbers with a <NUMBER> token or all names with a <NAME> token

Q68 How does the encoder-decoder structure work for language modelling?

The encoder-decoder structure is a deep learning model architecture responsible for several state of the art solutions, including Machine Translation.

The input sequence is passed to the encoder where it is transformed to a fixed-dimensional vector representation using a neural network. The transformed input is then decoded using another neural network. Then, these outputs undergo another transformation and a softmax layer. The final output is a vector of probabilities over the vocabularies. Meaningful information is extracted based on these probabilities.

Q69 What are attention mechanisms and why do we use them?

This was a followup to the encoder-decoder question. Only the output from the last time step is passed to the decoder, resulting in a loss of information learned at previous time steps. This information loss is compounded for longer text sequences with more time steps.

Attention mechanisms are a function of the hidden weights at each time step. When we use attention in encoder-decoder networks, the fixed-dimensional vector passed to the decoder becomes a function of all vectors outputted in the intermediary steps.

Two commonly used attention mechanisms are additive attention and multiplicative attention. As the names suggest, additive attention is a weighted sum while multiplicative attention is a weighted multiplier of the hidden weights. During the training process, the model also learns weights for the attention mechanisms to recognize the relative importance of each time step.

Q70 How would you implement an NLP system as a service, and what are some pitfalls you might face in production?

This is less of a NLP question than a question for productionizing machine learning models. There are however certain intricacies to NLP models.

..

Without diving too much into the productionization aspect, an ideal Machine Learning service will have:

● endpoint(s) that other business systems can use to make inference

● a feedback mechanism for validating model predictions

● a database to store predictions and ground truths from the feedback

● a workflow orchestrator which will (upon some signal) re-train and load the new model for serving based on the records from the database + any prior training data

● some form of model version control to facilitate rollbacks in case of bad deployments

● post-production accuracy and error monitoring

Q71 How can we handle misspellings for text input?

By using word embeddings trained over a large corpus (for instance, an extensive web scrape of billions of words), the model vocabulary would include common misspellings by design. The model can then learn the relationship between misspelled and correctly spelled words to

recognize their semantic similarity.
We can also preprocess the input to prevent misspellings. Terms not found in the model vocabulary can be mapped to the "closest" vocabulary term using:
● edit distance between strings
● phonetic distance between word pronunciations
● keyword distance to catch common typos

Q72 Which of the following models can perform tweet classification with regards to context mentioned above?

A) Naive Bayes
B) SVM
C) None of the above
Solution: (C)
Since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

Q73 You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, in regards to document term matrix?

1. Removal of stopwords from the data will affect the dimensionality of data
2. Normalization of words in the data will reduce the dimensionality of data
..
3. Converting all the words in lowercase will not affect the dimensionality of the data

A) Only 1
B) Only 2
C) Only 3
D) 1 and 2
E) 2 and 3
F) 1, 2 and 3
Solution: (D)
Choices A and B are correct because stopword removal will decrease the number of features in the matrix, normalization of words will also reduce redundant features, and, converting all words to lowercase will also decrease the dimensionality.

Q74 Which of the following features can be used for accuracy improvement of a classification model?

A) Frequency count of terms
B) Vector Notation of sentence
C) Part of Speech Tag
D) Dependency Grammar
E) All of these
Solution: (E)
All of the techniques can be used for the purpose of engineering features in a model.

Q75 What percentage of the total statements are correct with regards to Topic Modeling?

1. It is a supervised learning technique
2. LDA (Linear Discriminant Analysis) can be used to perform topic modeling
3. Selection of number of topics in a model does not depend on the size of data
4. Number of topic terms are directly proportional to size of the data

A) 0
B) 25
C) 50
D) 75
E) 100
Solution: (A)

LDA is unsupervised learning model, LDA is latent Dirichlet allocation, not Linear discriminant analysis. Selection of the number of topics is directly proportional to the size of the data, while number of topic terms is not directly proportional to the size of the data. Hence none of the statements are correct.

Q76 In Latent Dirichlet Allocation model for text classification purposes, what does alpha and beta hyperparameter represent-

A) Alpha: number of topics within documents, beta: number of terms within topics False

B) Alpha: density of terms generated within topics, beta: density of topics generated within terms False

C) Alpha: number of topics within documents, beta: number of terms within topics False

D) Alpha: density of topics generated within documents, beta: density of terms generated within topics True

Solution: (D)

Option D is correct

Q77 What is the problem with ReLu?

● Exploding gradient(Solved by gradient clipping)

● Dying ReLu — No learning if the activation is 0 (Solved by parametric relu)

● Mean and variance of activations is not 0 and 1.(Partially solved by subtracting around 0.5 from activation. Better explained in fastai videos)

Q78 What is the difference between learning latent features using SVD and getting embedding vectors using deep network?

SVD uses linear combination of inputs while a neural network uses nonlinear combination.

Q79 What is the information in the hidden and cell state of LSTM?

Hidden stores all the information till that time step and cell state stores particular information that might be needed in the future time step.

Number of parameters in an LSTM model with bias

$4(mh+h^2+h)$ where m is input vectors size and h is output vectors size a.k.a. hidden

The point to see here is that mh dictates the model size as m>>h. Hence it's important to have a small vocab.

Time complexity of LSTM

$seq\_length*hidden^2$

Time complexity of transfomer

$seq\_length^2*hidden$

When hidden size is more than the seq_length(which is normally the case), transfomer is faster than LSTM.

Q80 When is self-attention not faster than recurrent layers?

When the sequence length is greater than the representation dimensions. This is rare.

Q81 What is the benefit of learning rate warm-up?

..

Learning rate warm-up is a learning rate schedule where you have low (or lower) learning rate at the beginning of training to avoid divergence due to unreliable gradients at the beginning. As the model becomes more stable, the learning rate would increase to speed up convergence.

Q82 What's the difference between hard and soft parameter sharing in multi-task learning?

Hard sharing is where we train for all the task at the same time and update our weights using all the losses whereas soft sharing is where we train for one task at a time.

Q83 What's the difference between BatchNorm and LayerNorm?

BatchNorm computes the mean and variance at each layer for every minibatch whereas LayerNorm computes the mean and variance for every sample for each layer independently. Batch normalisation allows you to set higher learning rates, increasing speed of training as it reduces the unstability of initial starting weights.

Q84 Difference between BatchNorm and LayerNorm?

BatchNorm — Compute the mean and var at each layer for every minibatch
LayerNorm — Compute the mean and var for every single sample for each layer independently

Q85 Why does the transformer block have LayerNorm instead of BatchNorm?

Looking at the advantages of LayerNorm, it is robust to batch size and works better as it works at the sample level and not batch level.

Q86 What changes would you make to your deep learning code if you knew there are errors in your training data?

We can do label smoothening where the smoothening value is based on % error. If any particular class has known error, we can also use class weights to modify the loss.

Q87 What are the tricks used in ULMFiT? (Not a great questions but checks the awareness)

- LM tuning with task text
- Weight dropout
- Discriminative learning rates for layers
- Gradual unfreezing of layers
- Slanted triangular learning rate schedule

This can be followed up with a question on explaining how they help.

Q88 Tell me a language model which doesn't use dropout

ALBERT v2 — This throws a light on the fact that a lot of assumptions we take for granted are not necessarily true. The regularisation effect of parameter sharing in ALBERT is so strong that dropouts are not needed. (ALBERT v1 had dropouts.)

Q89 What are the differences between GPT and GPT-2? (From Lilian Weng)

..

- Layer normalization was moved to the input of each sub-block, similar to a residual unit of type "building block" (differently from the original type "bottleneck", it has batch normalization applied before weight layers).
- An additional layer normalization was added after the final self-attention block.
- A modified initialization was constructed as a function of the model depth.
- The weights of residual layers were initially scaled by a factor of 1/√n where n is the number of residual layers.
- Use larger vocabulary size and context size.

Q90 What are the differences between GPT and BERT?

- GPT is not bidirectional and has no concept of masking
- BERT adds next sentence prediction task in training and so it also has a segment embedding

Q91 What are the differences between BERT and ALBERT v2?

- Embedding matrix factorisation(helps in reducing no. of parameters)
- No dropout
- Parameter sharing(helps in reducing no. of parameters and regularisation)

Q92 How does parameter sharing in ALBERT affect the training and inference time?

No effect. Parameter sharing just decreases the number of parameters.

Q93 How would you reduce the inference time of a trained NN model?

- Serve on GPU/TPU/FPGA
- 16 bit quantisation and served on GPU with fp16 support
- Pruning to reduce parameters
- Knowledge distillation (To a smaller transformer model or simple neural network)
- Hierarchical softmax/Adaptive softmax
- You can also cache results as explained here.

Q94 Would you use BPE with classical models?

Of course! BPE is a smart tokeniser and it can help us get a smaller vocabulary which can help us find a model with less parameters.

Q95 How would you make an arxiv papers search engine? (I was asked — How would you make a plagiarism detector?)

Get top k results with TF-IDF similarity and then rank results with
● semantic encoding + cosine similarity
● a model trained for ranking
..
Q96 Get top k results with TF-IDF similarity and then rank results with
● semantic encoding + cosine similarity
● a model trained for ranking
Q97 How would you make a sentiment classifier?
This is a trick question. The interviewee can say all things such as using transfer learning and latest models but they need to talk about having a neutral class too otherwise you can have really good accuracy/f1 and still, the model will classify everything into positive or negative.
The truth is that a lot of news is neutral and so the training needs to have this class. The interviewee should also talk about how he will create a dataset and his training strategies like the selection of language model, language model fine-tuning and using various datasets for multitask learning.
Q98 What is the difference between regular expression and regular grammar?
A regular expression is the representation of natural language in the form of mathematical expressions containing a character sequence. On the other hand, regular grammar is the generator of natural language, defining a set of defined rules and syntax which the strings in the natural language must follow.
Q99 Why should we use Batch Normalization?
Once the interviewer has asked you about the fundamentals of deep learning architectures, they would move on to the key topic of improving your deep learning model's performance.
Batch Normalization is one of the techniques used for reducing the training time of our deep learning algorithm. Just like normalizing our input helps improve our logistic regression model, we can normalize the activations of the hidden layers in our deep learning model as well:
Q100 How is backpropagation different in RNN compared to ANN?
In Recurrent Neural Networks, we have an additional loop at each node:
This loop essentially includes a time component into the network as well. This helps in capturing sequential information from the data, which could not be possible in a generic artificial neural network.
This is why the backpropagation in RNN is called Backpropagation through Time, as in backpropagation at each time step

1.What is Numpy?

- Ans: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array obje

2. Why NumPy is used in Python?

- Ans: NumPy is a package in Python used for Scientific Computing. NumPy package is used to perform different operatio

3. How to Install Numpy in Windows?

- Ans:

Step 1: Download Python for Windows 10/8/7. First, download the Python executable binaries on your Windows system

Step 2: Run the Python executable installer. …

Step 3: Install pip on Windows 10/8/7. …

Step 4: Install Numpy in Python using pip on Windows 10/8/7.

- Installation Process of Numpy..

step1: Open the terminal

step2: type pip install numpy

import numpy as np

4. how to create 1D Array ?

```
num=[1,2,3]

num = np.array(num)

print("1d array : ",num)
```

5. How to create 2D Array ?

```
num2=[[1,2,3],[4,5,6]]

num2 = np.array(num2)

print("\n2d array : \n",num2)
```

6. how to create 3D Array or ND Array ?

```
num3=[[[1,2,3],[4,5,6],[7,8,9]]]

num3 = np.array(num3)

print("\n3d array : \n",num3)
```

7. how to identified datatyp for numpy array?

```
print("\n data type num 3 ", num3.dtype)
```

8. Print 1D array with 5 zeros.

```
arr = np.zeros(5)

print("single array:",arr)
```

9. print zeros with 2 rows and 3 columns ?

```
arr2 = np.zeros((2,3))

print("Array with 2 rows and 3 cols : \n", arr2)
```

10. Create array with eye() - diagonal values ?

```
arr3 = np.eye(4)

print("diaglonal values : \n", arr3)
```

11. use of diag() square matrix ?

```
arr3 = np.diag([1,2,3,4])
print("square matrix \n",arr3)
```

12. Print Range Between 1 To 15 and show 4 integers random numbers

```
rand_arr = np.random.randint(1,15,4)
print("\n random number from 1 to 15 ", rand_arr)
```

13. describe the example of seed() function? and how to use it ? why seed()?

seed() is used in the process of generating the same sequences of random numbers on a constant basis and use the same number at

```
# How many ever times we execute the above code, it will get us the same numbers that are picked randomly for the first time
np.random.seed(123)
rand_arr4 = np.random.randint(1,100,20)
print("seed() showing same number only : ", rand_arr4)
```

14. Print first position, last position and 2nd and 3rd position

```
num = np.array([5,15,25,35])
print("first position : ",num[0]) #5
print("third position : ",num[2]) #25
```

15. if we don't know last number of position how to show it by pragmatically?

```
num = np.array([5,15,25,35])
print("last indexing done by -1 position : ",num[-1])
```

16. create a matrix 3 * 3 with value ranging from 0 to 8

```
arr = np.arange(0,9).reshape(3,3)
print(arr)
```

17. create matrix 2 * 2 with value ranging from 1 to 3

```
arr = np.arange(0,4).reshape(2,2)
print(arr)
```

18.create matrix 2 * 2 with value ranging from 1 to 4

```
arr = np.arange(1,5).reshape(2,2)
print(arr)
```

19. print random number from 0 to 1

```
a = np.random.rand()
```

```
print("random number from 0 to 1 ", a)
```

20. print size of num array

```
num = np.array([5,15,25,35])
print("\n size \n",num.size) #100
```

21. create an array of 20 linearly spaced point between 0 to 1

```
num_line = np.linspace(0,1,20)
print(num_line)
```

22. Use of sort() function ?

```
num_arr = np.array([[5,6,2],[9,8,1]])
print("sorting of numpy array : \n",np.sort(num_arr))
```

23. How to get the positions where elements of two arrays match?

```
a = np.array([1,2,3,2,3,4,3,4,5,6])
b = np.array([7,2,10,2,7,4,9,4,9,8])
np.where(a == b)
```

24. Create a Numpy array filled with all zeros

```
a = np.zeros(3, dtype = int)
print("Matrix a : \n", a)
b = np.zeros([3, 3], dtype = int)
print("\nMatrix b : \n", b)
```

25. Check whether a Numpy array contains a specified row

```
arr = np.array([[1, 2, 3, 4, 5],
                [6, 7, 8, 9, 10],
                [11, 12, 13, 14, 15],
                [16, 17, 18, 19, 20]])
# view the array
arr
## tolist(): used to convert the data elements of an array into a list.
arr.tolist()
```

checking whether the given list is present in the above array.

```
print([1, 2, 3, 4, 5] in arr.tolist())

print([16, 17, 20, 19, 18] in arr.tolist())

print([3, 2, 5, -4, 5] in arr.tolist())

print([11, 12, 13, 14, 15] in arr.tolist())
```

26. Find the number of occurrences of a sequence in a NumPy array

```
rr = np.array([[2, 8, 9, 4],
               [9, 4, 9, 4],
               [4, 5, 9, 7],
               [2, 9, 4, 3]])



output = repr(arr).count("9, 4")

print(output)
```

27.Write a NumPy program to test element-wise for NaN of a given array.

```
a = np.array([1, 0, np.nan, np.inf])

print("Original array")

print(a)

print("Test element-wise for NaN:")

print(np.isnan(a))
```

28.Write a NumPy program to test whether none of the elements of a given array

```
x = np.array([1, 2, 3, 4])

print("Original array:")

print(x)

print("Test if none of the elements of the said array is zero:")

print(np.all(x))

x = np.array([0, 1, 2, 3])

print("Original array:")

print(x)

print("Test if none of the elements of the said array is zero:")

print(np.all(x))
```

29. Write a NumPy program to convert a list of numeric value into a one-dimensional NumPy array.

```
l = [12.23, 13.32, 100, 36.32]
print("Original List:",l)
a = np.array(l)
print("One-dimensional NumPy array: ",a)
```

30. Write a NumPy program to create a null vector of size 10 and update sixth value to 11.

```
x = np.zeros(10)
print(x)
print("\n Update sixth value to 11")
x[6] = 11
print(x)
```

31. Write a NumPy program to create a 5x5 array with random values and find the minimum and maximum values.

```
x = np.random.random((5,5))
print("Original Array:")
print(x)
xmin, xmax = x.min(), x.max()
print("\n")
print("Minimum and Maximum Values:")
print(xmin, xmax)
```

31. Write a NumPy program to compute the multiplication of two given matrixes.

Sample Matrix:

[[1, 0], [0, 1]]

[[1, 2], [3, 4]]

```
p = [[1, 0], [0, 1]]
q = [[1, 2], [3, 4]]
print("original matrix:")
print(p)
print(q)
result1 = np.dot(p, q)
print("\n")
```

```
print("Result of the said matrix multiplication:")
print(result1)
```

33.Create a two-dimensional array with the flattened input as a diagonal

```
ar1 = np.diagflat([[1,2], [3,4]])
ar1
```

34. Write a NumPy program to compute the sum of the diagonal element of a given array.

```
m = np.arange(6).reshape(3, 2)
print("Original matrix:")
print(m)
# trace() will return the sum of diagonal elements
result =  np.trace(m)
print("\n")
print("Condition number of the said matrix:")
print(result)
a = np.arange(24).reshape((2,2,2,3))
print(a)
print("\n")
np.trace(a)
a = np.arange(8).reshape((2,2,2))
print(a)
print("\n")
np.trace(a)
```

35. Use of around() function: returns a decimal value rounded to a desired position of the decimal

```
arr = np.array([12.202, 90.23120, 123.020, 23.202])
print("printing the original array values:",end = " ")
print(arr)
print("\n")
print("Array values rounded off to 2 decimal position",np.around(arr, 2))
print("Array values rounded off to -1 decimal position",np.around(arr, -1))
```

Basic Statistics and Distributions:

1) **What is the difference between data analysis and machine learning?**

Data analysis requires strong knowledge of coding and basic knowledge of statistics.

Machine learning, on the other hand, requires basic knowledge of coding and strong knowledge of statistics and business.

2) **What is big data?**

Big data has 5 major components – volume (size of data), velocity (inflow of data) and variety (types of data), veracity and value.

Big data causes "overloads"

3) **What are the four main things we should know before studying data analysis?**

Descriptive statistics

Inferential statistics

Distributions (normal distribution / sampling distribution)

Hypothesis testing

4) **What is the difference between inferential statistics and descriptive statistics?**

Descriptive statistics – provides exact and accurate information.

Inferential statistics – provides information of a sample and we need inferential statistics to reach a conclusion about the population.

5) **What is the difference between population and sample in inferential statistics?**

From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

From the sample we calculate the statistics.

From the sample statistics we conclude about the population.

6) **What are descriptive statistics?**

Descriptive statistic is used to describe the data (data properties)

5-number summary is the most commonly used descriptive statistics

7) **Most common characteristics used in descriptive statistics?**

Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.

Mean – average of all the numbers

Median – the number in the middle

Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.

Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.

Range = Max – Min

InterQuartile Range (IQR) = Q3 – Q1

Standard Deviation (σ) = √(∑(x-μ)2 / n)

Variance = σ2

Shape – the shape of the data can be symmetric or skewed

Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Left skewed – the left tail is longer than the right side

Right skewed – the right tail is longer than the left side

Outlier – An outlier is an abnormal value

Keep the outlier based on judgement

Remove the outlier based on judgement

8) What is quantitative data and qualitative data?

Quantitative data is also known as numeric data

Qualitative data is also known as categorical data

9) How to calculate range and interquartile range?

IQR = Q3 – Q1

Where, Q3 is the third quartile (75 percentile)

Where, Q1 is the first quartile (25 percentile)

10) Why do we need a 5-number summary?

Low extreme (minimum)

Lower quartile (Q1)

Median

Upper quartile (Q3)

Upper extreme (maximum)

11) What is the benefit of using box plots?

Shows the 5-number summary pictorially

Can be used to find outliers and compare group of histograms

12) What is the meaning of standard deviation?

It represents how far are the data points from the mean

(σ) = √(∑(x-μ)2 / n)

Variance is the square of standard deviation

13. What is the left skewed distribution and right skewed distribution?

Left skewed

The left tail is longer than the right side

Mean < median < mode

Right skewed

The right tail is longer than the right side

Mode < median < mean

14. What does symmetric distribution mean?

The part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Few examples are – uniform distribution, binomial distribution, normal distribution

15. What is the relationship between mean and median in normal distribution?

In the normal distribution mean is equal to median

16. What does it mean by bell curve distribution and Gaussian distribution?

Normal distribution is called bell curve distribution / Gaussian distribution

It is called bell curve because it has the shape of a bell

It is called Gaussian distribution as it is named after Carl Gauss

17. How to convert normal distribution to standard normal distribution?

Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula

$X \text{ (standardized)} = (x-\mu) / \sigma$

18. What is an outlier?

An outlier is an abnormal value (It is at an abnormal distance from rest of the data points).

19. Mention one method to find outliers?

Shows the 5-number summary can be used to identify the outlier

Widely used – Any data point that lies outside the 1.5 * IQR

Lower bound = Q1 – (1.5 * IQR)

Upper bound = Q3 + (1.5 * IQR)

20. What can I do with outliers?

Generally, we first check the performance of model with outliers, then we impute them and check the model performance. And even before that check first how much percent of data is outliers. Based on that you can take decision.

Remove outlier

When we know the data-point is wrong (negative age of a person)

When we have lots of data

We should provide two analyses. One with outliers and another without outliers.

Keep outlier

When there are lot of outliers (skewed data)

When results are critical

When outliers have meaning (fraud data)

21. What is the difference between population parameters and sample statistics?

Population parameters are:

Mean = μ

Standard deviation = σ

Sample statistics are:

Mean = x (bar)

Standard deviation = s

22. Why do we need sample statistics?

Population parameters are usually unknown hence we need sample statistics.

23. How to find the mean length of all fishes in the sea?

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

24. What are the effects of the width of the confidence interval?

Confidence interval is used for decision making

As the confidence level increases the width of the confidence interval also increases

As the width of the confidence interval increases, we tend to get useless information also.

Useless information – wide CI

High risk – narrow CI

25. Mention the relationship between standard error and margin of error?

As the standard error increases the margin of error also increases

26. Mention the relationship between confidence interval and margin of error?

As the confidence level increases the margin of error also increases

27. What is the proportion of confidence intervals that will not contain the population parameter?

Alpha is the portion of confidence interval that will not contain the population parameter

$\alpha = 1 - CL$

28. What is the difference between 95% confidence level and 99% confidence level?

The confidence interval increases as me move from 95% confidence level to 99% confidence level

29. What do you mean by degree of freedom?

DF is defined as the number of options we have

DF is used with t-distribution and not with Z-distribution

For a series, DF = n-1 (where n is the number of observations in the series)

30. What do you think if DF is more than 30?

As DF increases the t-distribution reaches closer to the normal distribution

At low DF, we have fat tails

If DF > 30, then t-distribution is as good as normal distribution

31. When to use t distribution and when to use z distribution?

The following conditions must be satisfied to use Z-distribution

Do we know the population standard deviation?

Is the sample size > 30?

CI = x (bar) – $Z^*\sigma/\sqrt{n}$ to x (bar) + $Z^*\sigma/\sqrt{n}$

Else we should use t-distribution

CI = x (bar) – t*s/√n to x (bar) + t*s/√n

32. What is H0 and H1? What is H0 and H1 for two-tail test?

H0 is known as null hypothesis. It is the normal case / default case.

For one tail test $x <= \mu$

For two-tail test $x = \mu$

H1 is known as alternate hypothesis. It is the other case.

For one tail test $x > \mu$

For two-tail test $x <> \mu$

33. What is p-value in hypothesis testing?

In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true.

If the p-value is more than the critical value, then we fail to reject the H0

If p-value = 0.015 (critical value = 0.05) – strong evidence

If p-value = 0.055 (critical value = 0.05) – weak evidence

If the p-value is less than the critical value, then we reject the H0

If p-value = 0.055 (critical value = 0.05) – weak evidence

If p-value = 0.005 (critical value = 0.05) – strong evidence

34. How to calculate p-value using manual method?

Find H0 and H1

Find n, x(bar) and s

Find DF for t-distribution

Find the type of distribution – t or z distribution

Find t or z value (using the look-up table)

Compute the p-value to critical value

35. How to calculate p-value using EXCEL?

Go to Data tab

Click on Data Analysis

Select Descriptive Statistics

Choose the column

Select summary statistics and confidence level (0.95)

36. What do we mean by – making decision based on comparing p-value with significance level?

If the p-value is more than then critical value, then we fail to reject the H0

If the p-value is less than the critical value, then we reject the H0

37. What is the difference between one tail and two tail hypothesis testing?

2-tail test: Critical region is on both sides of the distribution

H0: $x = \mu$

H1: $x <> \mu$

1-tail test: Critical region is on one side of the distribution

H1: $x <= \mu$

H1: $x > \mu$

38. What do you think of the tail (one tail or two tail) if H0 is equal to one value only?

It is a two-tail test

39. What is the critical value in one tail or two-tail test?

Critical value in 1-tail = alpha

Critical value in 2-tail = alpha / 2

40. Why is the t-value same for 90% two tail and 95% one tail test?

P-value of 1-tail = P-value of 2-tail / 2

It is because in two tail there are 2 critical regions

41. What is the central limit theorem? Why it is important?

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

To give an example, you would take a sample from a data set and calculate the mean of that sample. Once repeated multiple times, you would plot all your means and their frequencies onto a graph and see that a bell curve, also known as a normal distribution, has been created. The mean of this distribution will closely resemble that of the original data. The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.

42. What is observational and experimental data in Statistics?

Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

43. What is meant by mean imputation for missing data?

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

Mean imputation is not bad. It depends how much percentage of data is missing. If more than 50% is missing than mean imputation will not make sense as we will not have variance in data.

44. What is an outlier? How can outliers be determined in a dataset?

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

Outliers are determined by using two methods:

Standard deviation/z-score

Interquartile range (IQR)

45. How is missing data handled in statistics?

There are many ways to handle missing data in Statistics:

- Prediction of the missing values
- Assignment of individual (unique) values
- Deletion of rows, which have the missing data
- Mean imputation or median imputation
- Using random forests, which support the missing values

46. What are the types of selection bias in statistics?

There are many types of selection bias as shown below:

Observer selection

Attrition

Protopathic bias

Time intervals

Sampling bias

47. What type of data does not have a log-normal distribution or a Gaussian distribution?

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

48. What is the meaning of the five-number summary in Statistics?

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

Low extreme (Min)

First quartile (Q1)

Median

Upper quartile (Q3)

High extreme (Max)

49. What is the meaning of standard deviation?

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

50 What is a bell-curve distribution?

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

51. What is skewness?

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

52. What is kurtosis?

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

53. What is correlation?

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

54. What are left-skewed and right-skewed distributions?

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

55. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

56.What is the relationship between the confidence level and the significance level in statistics?

The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

Significance level = 1 − Confidence level

57.What is the relationship between mean and median in a normal distribution?

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset's mean and median.

58.What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

The lower quartile (Q1) is the 25th percentile.

The middle quartile (Q2), also called the median, is the 50th percentile.

The upper quartile (Q3) is the 75th percentile.

59.How do the standard error and the margin of error relate?

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

60. What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population's mean.

61. What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

62. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

63. What are the types of biases that we encounter while sampling?

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis.

There are three types of biases:

The selection bias

The survivorship bias

The undercoverage bias

64. What are some of the techniques to reduce underfitting and overfitting during model training?

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

<u>For reducing underfitting:</u>

Increase model complexity

Increase the number of features

Remove noise from the data

Increase the number of training epochs

<u>For reducing overfitting:</u>

Increase training data

Stop early while training

Lasso regularization

Use random dropouts

65. What is the assumption of normality?

The assumption of normality is the sampling distribution is normal and centers around the population parameter, according to the central limit theorem.

66. What is the difference between type 1 error and type 2 error?

A type 1 error is when you incorrectly reject a true null hypothesis. It's also called a false positive.

A type 2 error is when you don't reject a false null hypothesis. It's also called a false negative.

68. How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. *Variance* is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

69. A data set is given to you and it has missing values which spread along 1 standard deviation from the mean. How much of the data would remain untouched?

It is given that the data is spread across mean that is the data is spread across an average. So, we can presume that it is a normal distribution. In a normal distribution, about 68% of data lies in 1 standard deviation from averages like mean, mode or median. That means about 32% of the data remains uninfluenced by missing values.

70. Is a high variance in data good or bad?

Higher variance directly means that the data spread is big and the feature has a variety of data. Usually, high variance in a feature is seen as not so good quality.

71. If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

72. Explain the difference between Normalization and Standardization.

Normalization and Standardization are the two very popular methods used for feature scaling.

Normalization refers to re-scaling the values to fit into a range of [0,1]. Standardization refers to re-scaling data to have a mean of 0 and a standard deviation of 1 (Unit variance). Normalization is useful when all parameters need to have the identical positive scale however the outliers from the data set are lost. Hence, standardization is recommended for most applications.

73. What is the meaning of covariance?

Covariance is the measure of indication when two items vary together in a cycle. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

74. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 2.

75. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

True, a normal curve will have the area under unity and the symmetry around zero in any distribution. Here, all of the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.

76. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

77. A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?

If the oranges are increased by one, there will be an increase of 2 apples since the equation is:

$y = 100 + 2x.$

78. What are the examples of symmetric distribution?

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

Uniform distribution

Binomial distribution

Normal distribution

79. What Is Null Hypothesis?

The null hypothesis (denote by H0) is a statement about the value of a population parameter (such as mean), and it must contain the condition of equality and must be written with the symbol =, ≤, or ≤.

80. **What Is Sampling?**

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern.

81. What Are Sampling Methods?

There are four sampling methods:

Simple Random (purely random),

Systematic (eg. every kth member of population),

Cluster (population divided into groups or clusters)

Stratified (divided by exclusive groups or strata, sample from each group) samplings.

82. What are the measures of Dispersion / Spread?

We measure spread using range, interquartile range, variance and standard deviation.

83. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). Sensitivity is nothing but "Predicted True events/ Total events". True events here are the events which were true and model also predicted them as true. Calculation of seasonality is pretty straightforward. Seasonality = (True Positives) / (Positives in Actual Dependent Variable)

84. What is Re-sampling?

Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.

85. Why Is Re-sampling Done?

Resampling is done in any of these cases:

• Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points

• Substituting labels on data points when performing significance tests

• Validating models by using random subsets (bootstrapping, cross-validation)

86. What are the Non-probability data sampling methods?

Non-probability data sampling methods include:

• Convenience sampling: Data is collected from an easily accessible and available group.

• Consecutive sampling: Data is collected from every subject that meets the criteria until the predetermined sample size is met.

• Purposive or judgmental sampling: The researcher selects the data to sample based on predefined criteria.

• Quota sampling: The researcher ensures equal representation within the sample for all subgroups in the data set or population (random sampling is not used).

87. How does data cleaning play a vital role in the analysis?

Data cleaning can help in analysis because:

• Cleaning data from multiple sources helps transform it into a format that data analysts or data scientists can work with.

• Data Cleaning helps increase the accuracy of the model in machine learning.

• It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.

• It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

88. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

P(Having two girls given one girl) = ½

89. What is probability density function?

The Probability Density Function (PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution or just a probability function.

90. What is Hypothesis Testing?

Hypothesis testing in statistics refers to analyzing an assumption about a population parameter. It is used to make an educated guess about an assumption using statistics. With the use of sample data, hypothesis testing makes an assumption about how true the assumption is for the entire population from where the sample is being taken.

91. What are the 4 stages of hypothesis testing?

All hypotheses are tested using a four-step process:

The first step is for the analyst to state the two hypotheses so that only one can be right.

The next step is to formulate an analysis plan, which outlines how the data will be evaluated.

The third step is to carry out the plan and physically analyse the sample data.

The fourth and final step is to analyse the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

Why do statisticians use interquartile range?

When measuring variability, statisticians prefer using the interquartile range instead of the full data range because extreme values and outliers affect it less. Typically, use the IQR with a measure of central tendency, such as the median, to understand your data's centre and spread.

What is the Pearson correlation coefficient?

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

For eg: Up till a certain age, (in most cases) a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.

This approach is based on covariance and thus is the best method to measure the relationship between two variables.

What is Z-score?

Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean.

What are the types of data?

Data can be categorized as follows;

Qualitative data (Categorical)

- Nominal
- Ordinal

Quantitative data (Numerical)

- Discrete

- Continuous

What is nominal and ordinal data?

Categorical variables are broken down into nominal and ordinal data.

Nominal data (also known as nominal scale) is a classification of categorical variables.

Ordinal data is a kind of categorical data with a set order or scale to it. For example, ordinal data is said to have been collected when a responder inputs his/her financial happiness level on a scale of 1-10. In ordinal data, there is no standard scale on which the difference in each score is measured.

What is discrete and continuous data?

Numerical variables are classified into continuous and discrete data. Data that can only take on certain values are discrete data. These values do not have to be complete numbers, but they are values that are fixed. It only contains finite values, the subdivision of which is not possible. It includes only those values which are separate and can only be counted in whole numbers or integers which means that the data cannot be split into fractions or decimals.

Eg: The number of students in a class, the number of chocolates in a bag, the number of strings on the guitar, the number of fishes in the aquarium, etc.

Continuous data is the data that can be of any value. Over time, some continuous data can change. It may take any numeric value, within a potential value range of finite or infinite. The continuous data can be broken down into fractions or decimals i.e. according to measurement accuracy, it can be significantly subdivided into smaller sections.

Eg: Measurement of height and weight of a student, Daily temperature measurement of a place, Wind speed measured daily, etc.

What is univariate analysis?

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

Univariate analysis is conducted through several ways which are mostly descriptive in nature –

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

What is bivariate analysis?

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Bivariate analysis is conducted using –

• Correlation coefficients

• Regression analysis

What is multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

Commonly used multivariate analysis technique include –

• Factor Analysis

• Cluster Analysis

• Variance Analysis

• Discriminant Analysis

• Multidimensional Scaling

• Principal Component Analysis

• Redundancy Analysis

Python Scikit-Learn Cheat Sheet

If you are finding it hard to remember all the different commands to perform different operations in Scikit Learn then don't worry, you are not alone, it happens more often than you would think.

What is Scikit Learn?

Scikit-Learn or "sklearn" is a free, open source machine learning library for the Python programming language. It's simple yet efficient tool for data mining, Data analysis and Machine Learning. It features various machine learning algorithms and also supports Python's scientific and numerical libraries, that is, SciPy and NumPy respectively. Get back to learn Python for all other topics..

Import Convention

Before you can start using Scikit-learn, you need to remember that it is a Python library and you need to import it. To do that all you have to do is type the following command:

```
import sklearn
```

Preprocessing

The process of converting raw data set into a meaningful and clean data set is referred to as Preprocessing of data. This is a 'must- follow' technique before you can feed your data set to a machine learning algorithm. There are mainly three steps that you need to follow while preprocessing the data. The steps are listed below:

1. Data Loading:

You need your data in numeric form stored in numeric arrays. Following are the two ways you can load the data, you can also use some other numeric array to load your data.

Using NumPy :

```
import numpy as np

a=np.array([(1,2,3,4),(7,8,9,10)],dtype=int)

data = np.loadtxt('file_name.csv', delimiter=',')
```

Using Pandas :

```
import pandas as pd

df=pd.read_csv('file_name.csv',header=0)
```

2. Train-Test data:

The next step is to split your data in training data set and testing data set

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=0)
```

3. Data Preparation:

Standardization: It makes the training process well behaved improving the numerical condition of the optimization problems.

```
from sklearn.preprocessing import StandardScaler

get_names = df.columns

scaler = preprocessing.StandardScaler()

scaled_df = scaler.fit_transform(df)

scaled_df = pd.DataFrame(scaled_df, columns=get_names)
```

Normalization: It makes training less sensitive to the scale of features, also makes the data better conditioned for convergence.

```
from sklearn.preprocessing import Normalizer

pd.read_csv("File_name.csv")

x_array = np.array(df['Column1'])#Normalize Column1

normalized_X = preprocessing.normalize([x_array])
```

Working on a model

After making all the necessary transformation in our dataset, in order to make it algorithm-ready, we need to work on our model, that is, choosing a correct model or an algorithm that represents our dataset and will help us make the kind of predictions that we want from our chosen data set and then performing model fitting.

Model Choosing:

Supervised Learning Estimator:

Supervised learning, as the name suggests, is the kind of machine learning where we supervise the outcome by training the model with well labeled data, which means that some of the data in the dataset

will already be tagged with correct answers.

a. Linear Regression:

```
from sklearn.linear_model import LinearRegression
new_lr = LinearRegression(normalize=True)
```

b. Support Vector Machine:

```
from sklearn.svm import SVC
new_svc = SVC(kernel='linear')
```

c. Naive Bayes:

```
from sklearn.naive_bayes import GaussianNB
new_gnb = GaussianNB()
```

d. KNN:

```
from sklearn import neighbors
knn=neighbors.KNeighborsClassifier(n_neighbors=1)
```

Unsupervised Learning Estimator:

Unlike Supervised learning, unsupervised learning is where we train the model with non labeled data or non classified data and let the algorithm do all the work on that dataset without any assistance.

a. Principal Component Analysis (PCA):

```
from sklearn.decomposition import PCA
new_pca= PCA(n_components=0.95)
```

b. K Means:

```
from sklearn.cluster import KMeans

k_means = KMeans(n_clusters=5, random_state=0)
```

Model Fitting:

The goal of implementing model fitting is to learn how well a model will generalize when trained with a dataset similar to the dataset that the model was initially trained on. The more fitting model will produce more accurate outcomes.

Supervised:

```
new_ lr.fit(X, y)

knn.fit(X_train, y_train)

new_svc.fit(X_train, y_train)
```

unSupervised:

```
k_means.fit(X_train)

pca_model_fit = new_pca.fit_transform(X_train)
```

Post-Processing

After getting comfortable with our dataset and model, the next step is to finally follow the main goal of machine learning algorithms, that is, to forecast outcomes and make predictions.

Prediction:

Once you are done with choosing and fiting the model, you can make predictions on your dataset.

Supervised

Supervised:

```
y_predict = new_svc.predict(np.random.random((3,5)))

y_predict = new_lr.predict(X_test)

y_predict = knn.predict_proba(X_test)
```

Unsupervised:

```
y_pred = k_means.predict(X_test)
```

Evaluate Performance:

Evaluating the predictive performance of your model is necessary. There are multiple techniques in machine learning that can be used to organize classifiers and visualize their performance. Following are the said technologies.

Classification:

a. Confusion Matrix:

```
 from sklearn.metrics import confusion_matrix

print(confusion_matrix(y_test, y_pred))
```

b. Accuracy Score

```
knn.score(X_test, y_test)

 from sklearn.metrics import accuracy_score

accuracy_score(y_test, y_pred)
```

Regression:

a. Mean Absolute Error:

```
 from sklearn.metrics import mean_absolute_error

y_true = [3, -0.5, 2]

mean_absolute_error(y_true, y_predict)
```

b. Mean Squared Error:

```
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test, y_predict)
```

c. R^2 Score

```
from sklearn.metrics import r2_score r2_score(y_true, y_predict)
```

Clustering:

a. Homogeneity:

```
from sklearn.metrics import homogeneity_score
homogeneity_score(y_true, y_predict)
```

b. V-measure:

```
from sklearn.metrics import v_measure_score
metrics.v_measure_score(y_true, y_predict)
```

c. Cross-validation:

```
from sklearn.cross_validation import cross_val_score
print(cross_val_score(knn, X_train, y_train, cv=4))
print(cross_val_score(new_lr, X, y, cv=2))
```

Model Tuning:

This is the final step when implementing machine learning, before presenting the final outcomes. In Model tuning, models are parameterized so their behavior is tuned for a given problem. This is done by searching for the right set of parameters and we have mainly two ways of doing that:

Grid Search:

In Grid search, parameter tuning is done methodically and then it evaluates model for each set of parameter that is specified in a grid.

```
from sklearn.grid_search import GridSearchCV

params = {"n_neighbors": np.arange(1,3), "metric": ["euclidean", "cityblock"]}

grid = GridSearchCV(estimator=knn, param_grid=params)

grid.fit(X_train, y_train)

print(grid.best_score_)

print(grid.best_estimator_.n_neighbors)
```

Randomized Parameter Optimization:

In Randomised Search, random search is performed on a fixed set of parameters. The number of parameters that are used is given by n-iter.

```
from sklearn.grid_search import RandomizedSearchCV

params = {"n_neighbors": range(1,5), "weights": ["uniform", "distance"]}

rsearch = RandomizedSearchCV(estimator=knn, param_distributions=params,cv=4,n_iter=8, random_state=5)

rsearch.fit(X_train, y_train)

print(rsearch.best_score_)
```

Questions: Basic Statistics and Distributions:

13) What is the difference between data analysis and machine learning?

Data analysis requires strong knowledge of coding and basic knowledge of statistics.

Machine learning, on the other hand, requires basic knowledge of coding and strong knowledge of statistics and business.

14) What is big data?

Big data has 5 major components – volume (size of data), velocity (inflow of data) and variety (types of data), veracity and value.

Big data causes “overloads”

15) What are the four main things we should know before studying data analysis?

Descriptive statistics

Inferential statistics

Distributions (normal distribution / sampling distribution)

Hypothesis testing

16) What is the difference between inferential statistics and descriptive statistics?

Descriptive statistics – provides exact and accurate information.

Inferential statistics – provides information of a sample and we need inferential statistics to reach a conclusion about the population.

17) What is the difference between population and sample in inferential statistics?

From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

From the sample we calculate the statistics.

From the sample statistics we conclude about the population.

18) What are descriptive statistics?

Descriptive statistic is used to describe the data (data properties)

5-number summary is the most commonly used descriptive statistics

19) Most common characteristics used in descriptive statistics?

Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.

Mean – average of all the numbers

Median – the number in the middle

Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.

Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.

Range = Max – Min

InterQuartile Range (IQR) = Q3 – Q1

Standard Deviation (σ) = √(∑(x-μ)2 / n)

Variance = σ2

Shape – the shape of the data can be symmetric or skewed

Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Left skewed – the left tail is longer than the right side

Right skewed – the right tail is longer than the left side

Outlier – An outlier is an abnormal value

Keep the outlier based on judgement

Remove the outlier based on judgement

20) What is quantitative data and qualitative data?

Quantitative data is also known as numeric data

Qualitative data is also known as categorical data

21) How to calculate range and interquartile range?

IQR = Q3 – Q1

Where, Q3 is the third quartile (75 percentile)

Where, Q1 is the first quartile (25 percentile)

22) Why do we need a 5-number summary?

Low extreme (minimum)

Lower quartile (Q1)

Median

Upper quartile (Q3)

Upper extreme (maximum)

23) What is the benefit of using box plots?

Shows the 5-number summary pictorially

Can be used to find outliers and compare group of histograms

24) What is the meaning of standard deviation?

It represents how far are the data points from the mean

$(\sigma) = \sqrt{(\sum(x-\mu)2 / n)}$

Variance is the square of standard deviation

13. What is the left skewed distribution and right skewed distribution?

Left skewed

The left tail is longer than the right side

Mean < median < mode

Right skewed

The right tail is longer than the right side

Mode < median < mean

14. What does symmetric distribution mean?

The part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Few examples are – uniform distribution, binomial distribution, normal distribution

15. What is the relationship between mean and median in normal distribution?

In the normal distribution mean is equal to median

16. What does it mean by bell curve distribution and Gaussian distribution?

Normal distribution is called bell curve distribution / Gaussian distribution

It is called bell curve because it has the shape of a bell

It is called Gaussian distribution as it is named after Carl Gauss

17. How to convert normal distribution to standard normal distribution?

Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula

$X \text{ (standardized)} = (x-\mu) / \sigma$

18. What is an outlier?

An outlier is an abnormal value (It is at an abnormal distance from rest of the data points).

19. Mention one method to find outliers?

Shows the 5-number summary can be used to identify the outlier

Widely used – Any data point that lies outside the 1.5 * IQR

Lower bound = Q1 – (1.5 * IQR)

Upper bound = Q3 + (1.5 * IQR)

20. What can I do with outliers?

Generally, we first check the performance of model with outliers, then we impute them and check the model performance. And even before that check first how much percent of data is outliers. Based on that you can take decision.

Remove outlier

When we know the data-point is wrong (negative age of a person)

When we have lots of data

We should provide two analyses. One with outliers and another without outliers.

Keep outlier

When there are lot of outliers (skewed data)

When results are critical

When outliers have meaning (fraud data)

21. What is the difference between population parameters and sample statistics?

Population parameters are:

Mean = μ

Standard deviation = σ

Sample statistics are:

Mean = $\bar{x}$ (bar)

Standard deviation = s

22. Why do we need sample statistics?

Population parameters are usually unknown hence we need sample statistics.

23. How to find the mean length of all fishes in the sea?

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

24. What are the effects of the width of the confidence interval?

Confidence interval is used for decision making

As the confidence level increases the width of the confidence interval also increases

As the width of the confidence interval increases, we tend to get useless information also.

Useless information – wide CI

High risk – narrow CI

25. Mention the relationship between standard error and margin of error?

As the standard error increases the margin of error also increases

26. Mention the relationship between confidence interval and margin of error?

As the confidence level increases the margin of error also increases

27. What is the proportion of confidence intervals that will not contain the population parameter?

Alpha is the portion of confidence interval that will not contain the population parameter

$\alpha = 1 - CL$

28. What is the difference between 95% confidence level and 99% confidence level?

The confidence interval increases as I move from 95% confidence level to 99% confidence level

29. What do you mean by degree of freedom?

DF is defined as the number of options we have

DF is used with t-distribution and not with Z-distribution

For a series, DF = n-1 (where n is the number of observations in the series)

30. What do you think if DF is more than 30?

As DF increases the t-distribution reaches closer to the normal distribution

At low DF, we have fat tails

If DF > 30, then t-distribution is as good as normal distribution

31. When to use t distribution and when to use z distribution?

The following conditions must be satisfied to use Z-distribution

Do we know the population standard deviation?

Is the sample size > 30?

CI = x (bar) – Z*σ/√n to x (bar) + Z*σ/√n

Else we should use t-distribution

CI = x (bar) – t*s/√n to x (bar) + t*s/√n

32. What is H0 and H1? What is H0 and H1 for two-tail test?

H0 is known as null hypothesis. It is the normal case / default case.

For one tail test x <= μ

For two-tail test x = μ

H1 is known as alternate hypothesis. It is the other case.

For one tail test x > μ

For two-tail test x <> μ

33. What is p-value in hypothesis testing?

In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true.

If the p-value is more than the critical value, then we fail to reject the H0

If p-value = 0.015 (critical value = 0.05) – strong evidence

If p-value = 0.055 (critical value = 0.05) – weak evidence

If the p-value is less than the critical value, then we reject the H0

If p-value = 0.055 (critical value = 0.05) – weak evidence

If p-value = 0.005 (critical value = 0.05) – strong evidence

34. How to calculate p-value using manual method?

Find H0 and H1

Find n, x(bar) and s

Find DF for t-distribution

Find the type of distribution – t or z distribution

Find t or z value (using the look-up table)

Compute the p-value to critical value

35. How to calculate p-value using EXCEL?

Go to Data tab

Click on Data Analysis

Select Descriptive Statistics

Choose the column

Select summary statistics and confidence level (0.95)

36. What do we mean by – making decision based on comparing p-value with significance level?

If the p-value is more than then critical value, then we fail to reject the H0

If the p-value is less than the critical value, then we reject the H0

37. What is the difference between one tail and two tail hypothesis testing?

2-tail test: Critical region is on both sides of the distribution

H0: $x = \mu$

H1: $x <> \mu$

1-tail test: Critical region is on one side of the distribution

H1: $x <= \mu$

H1: $x > \mu$

38. What do you think of the tail (one tail or two tail) if H0 is equal to one value only?

It is a two-tail test

39. What is the critical value in one tail or two-tail test?

Critical value in 1-tail = alpha

Critical value in 2-tail = alpha / 2

40. Why is the t-value same for 90% two tail and 95% one tail test?

P-value of 1-tail = P-value of 2-tail / 2

It is because in two tail there are 2 critical regions

41. What is the central limit theorem? Why it is important?

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution

This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

To give an example, you would take a sample from a data set and calculate the mean of that sample. Once repeated multiple times, you would plot all your means and their frequencies onto a graph and see that a bell curve, also known as a normal distribution, has been created. The mean of this distribution will closely resemble that of the original data. The central limit theorem is important because it is used in hypothesis testing and als o to calculate confidence intervals.

42. What is observational and experimental data in Statistics?

Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

43. What is meant by mean imputation for missing data?

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

Mean imputation is not bad. It depends how much percentage of data is missing. If more than 50% is missing than mean imputation will not make sense as we will not have variance in data.

44. What is an outlier? How can outliers be determined in a dataset?

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

Outliers are determined by using two methods:

Standard deviation/z-score

Interquartile range (IQR)

45. How is missing data handled in statistics?

There are many ways to handle missing data in Statistics:

- Prediction of the missing values
- Assignment of individual (unique) values
- Deletion of rows, which have the missing data
- Mean imputation or median imputation
- Using random forests, which support the missing values

46. What are the types of selection bias in statistics?

There are many types of selection bias as shown below:

Observer selection

Attrition

Protopathic bias

Time intervals

Sampling bias

47. What type of data does not have a log-normal distribution or a Gaussian distribution?

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

48. What is the meaning of the five-number summary in Statistics?

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

Low extreme (Min)

First quartile (Q1)

Median

Upper quartile (Q3)

High extreme (Max)

49. What is the meaning of standard deviation?

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

50 What is a bell-curve distribution?

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

51. What is skewness?

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

52. What is kurtosis?

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

53. What is correlation?

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

54. What are left-skewed and right-skewed distributions?

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

55. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

56.What is the relationship between the confidence level and the significance level in statistics?

The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

Significance level = 1 − Confidence level

57.What is the relationship between mean and median in a normal distribution?

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset's mean and median.

58.What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

The lower quartile (Q1) is the 25th percentile.

The middle quartile (Q2), also called the median, is the 50th percentile.

The upper quartile (Q3) is the 75th percentile.

59.How do the standard error and the margin of error relate?

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

60. What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population's mean.

61. What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null

hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

62. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

63. What are the types of biases that we encounter while sampling?

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis.

There are three types of biases:

The selection bias

The survivorship bias

The undercoverage bias

64. What are some of the techniques to reduce underfitting and overfitting during model training?

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

<u>For reducing underfitting:</u>

Increase model complexity

Increase the number of features

Remove noise from the data

Increase the number of training epochs

For reducing overfitting:

Increase training data

Stop early while training

Lasso regularization

Use random dropouts

65. What is the assumption of normality?

The assumption of normality is the sampling distribution is normal and centers around the population parameter, according to the central limit theorem.

66. What is the difference between type 1 error and type 2 error?

A type 1 error is when you incorrectly reject a true null hypothesis. It's also called a false positive.

A type 2 error is when you don't reject a false null hypothesis. It's also called a false negative.

68. How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. *Variance* is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

69. A data set is given to you and it has missing values which spread along 1 standard deviation from the mean. How much of the data would remain untouched?

It is given that the data is spread across mean that is the data is spread across an average. So, we can presume that it is a normal distribution. In a normal distribution, about 68% of data lies in 1 standard deviation from averages like mean, mode or median. That means about 32% of the data remains uninfluenced by missing values.

70. Is a high variance in data good or bad?

Higher variance directly means that the data spread is big and the feature has a variety of data. Usually, high variance in a feature is seen as not so good quality.

71. If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

72. Explain the difference between Normalization and Standardization.

Normalization and Standardization are the two very popular methods used for feature scaling.

Normalization refers to re-scaling the values to fit into a range of [0,1]. Standardization refers to re-scaling data to have a mean of 0 and a standard deviation of 1 (Unit variance). Normalization is useful when all parameters need to have the identical positive scale however the outliers from the data set are lost. Hence, standardization is recommended for most applications.

73. What is the meaning of covariance?

Covariance is the measure of indication when two items vary together in a cycle. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

74. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 2.

75. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

True, a normal curve will have the area under unity and the symmetry around zero in any distribution. Here, all of the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.

76. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

77. A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?

If the oranges are increased by one, there will be an increase of 2 apples since the equation is:

y = 100 + 2x.

78. What are the examples of symmetric distribution?

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

Uniform distribution

Binomial distribution

Normal distribution

79. What Is Null Hypothesis?

The null hypothesis (denote by H0) is a statement about the value of a population parameter (such as mean), and it must contain the condition of equality and must be written with the symbol =, ≤, or ≤.

80. **What Is Sampling?**

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern.

81. What Are Sampling Methods?

There are four sampling methods:

- Simple Random (purely random),
- Systematic (eg. every kth member of population),
- Cluster (population divided into groups or clusters)
- Stratified (divided by exclusive groups or strata, sample from each group) samplings.

82. What are the measures of Dispersion / Spread?

We measure spread using range, interquartile range, variance and standard deviation.

83. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). Sensitivity is nothing but "Predicted True events/ Total events". True events here are the events which were true and model also predicted them as true. Calculation of seasonality is pretty straightforward. Seasonality = (True Positives) / (Positives in Actual Dependent Variable)

84. What is Re-sampling?

Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.

85. Why Is Re-sampling Done?

Resampling is done in any of these cases:

• Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points

• Substituting labels on data points when performing significance tests

• Validating models by using random subsets (bootstrapping, cross-validation)

86. What are the Non-probability data sampling methods?

Non-probability data sampling methods include:

• Convenience sampling: Data is collected from an easily accessible and available group.

• Consecutive sampling: Data is collected from every subject that meets the criteria until the predetermined sample size is met.

• Purposive or judgmental sampling: The researcher selects the data to sample based on predefined criteria.

• Quota sampling: The researcher ensures equal representation within the sample for all subgroups in the data set or population (random sampling is not used).

87. How does data cleaning play a vital role in the analysis?

Data cleaning can help in analysis because:

• Cleaning data from multiple sources helps transform it into a format that data analysts or data scientists can work with.

• Data Cleaning helps increase the accuracy of the model in machine learning.

• It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.

• It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

88. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

P(Having two girls given one girl) = ½

89. What is probability density function?

The Probability Density Function (PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution or just a probability function.

90. What is Hypothesis Testing?

Hypothesis testing in statistics refers to analyzing an assumption about a population parameter. It is used to make an educated guess about an assumption using statistics. With the use of sample data, hypothesis testing makes an assumption about how true the assumption is for the entire population from where the sample is being taken.

91. What are the 4 stages of hypothesis testing?

All hypotheses are tested using a four-step process:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyse the sample data.
4. The fourth and final step is to analyse the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

92. Why do statisticians use interquartile range?

When measuring variability, statisticians prefer using the interquartile range instead of the full data range because extreme values and outliers affect it less. Typically, use the IQR with a measure of central tendency, such as the median, to understand your data's centre and spread.

93. What is the Pearson correlation coefficient?

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

For eg: Up till a certain age, (in most cases) a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.

This approach is based on covariance and thus is the best method to measure the relationship between two variables.

94. What is Z-score?

Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean.

95. What are the types of data?

Data can be categorized as follows;

- Qualitative data (Categorical)
  - Nominal
  - Ordinal
- Quantitative data (Numerical)
  - Discrete
  - Continuous

96. What is nominal and ordinal data?

Categorical variables are broken down into nominal and ordinal data.

Nominal data (also known as nominal scale) is a classification of categorical variables.

Ordinal data is a kind of categorical data with a set order or scale to it. For example, ordinal data is said to have been collected when a responder inputs his/her financial happiness level on a scale of 1-10. In ordinal data, there is no standard scale on which the difference in each score is measured.

97. What is discrete and continuous data?

Numerical variables are classified into continuous and discrete data. Data that can only take on certain values are discrete data. These values do not have to be complete numbers, but they are values that are fixed. It only contains finite values, the subdivision of which is not possible. It includes only those values which are separate and can only be counted in whole numbers or integers which means that the data cannot be split into fractions or decimals.

Eg: The number of students in a class, the number of chocolates in a bag, the number of strings on the guitar, the number of fishes in the aquarium, etc.

Continuous data is the data that can be of any value. Over time, some continuous data can change. It may take any numeric value, within a potential value range of finite or infinite. The continuous data can be broken down into fractions or decimals i.e. according to measurement accuracy, it can be significantly subdivided into smaller sections.

Eg: Measurement of height and weight of a student, Daily temperature measurement of a place, Wind speed measured daily, etc.

98. What is univariate analysis?

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

Univariate analysis is conducted through several ways which are mostly descriptive in nature –

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

99. What is bivariate analysis?

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Bivariate analysis is conducted using –

- Correlation coefficients
- Regression analysis

100. What is multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

Commonly used multivariate analysis technique include –

- Factor Analysis
- Cluster Analysis
- Variance Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Principal Component Analysis
- Redundancy Analysis

Data Science Interview Questions and Answers: Basic to Technical

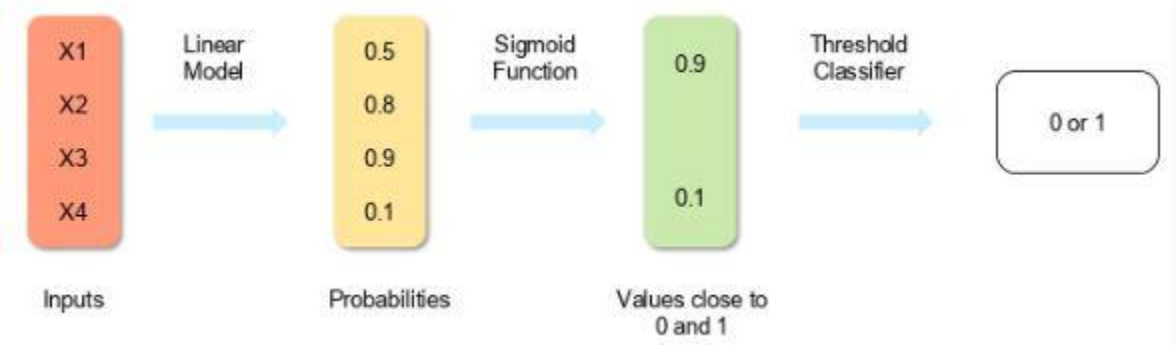
1. What are the differences between supervised and unsupervised learning?

Supervised Learning	Unsupervised Learning
Uses known and labeled data as input Supervised learning has a feedback mechanism The most commonly used supervised learning algorithms are decision trees, logistic regression, and support vector machine	Uses unlabeled data as input Unsupervised learning has no feedback mechanism The most commonly used unsupervised learning algorithms are k-means clustering, hierarchical clustering, and apriori algorithm

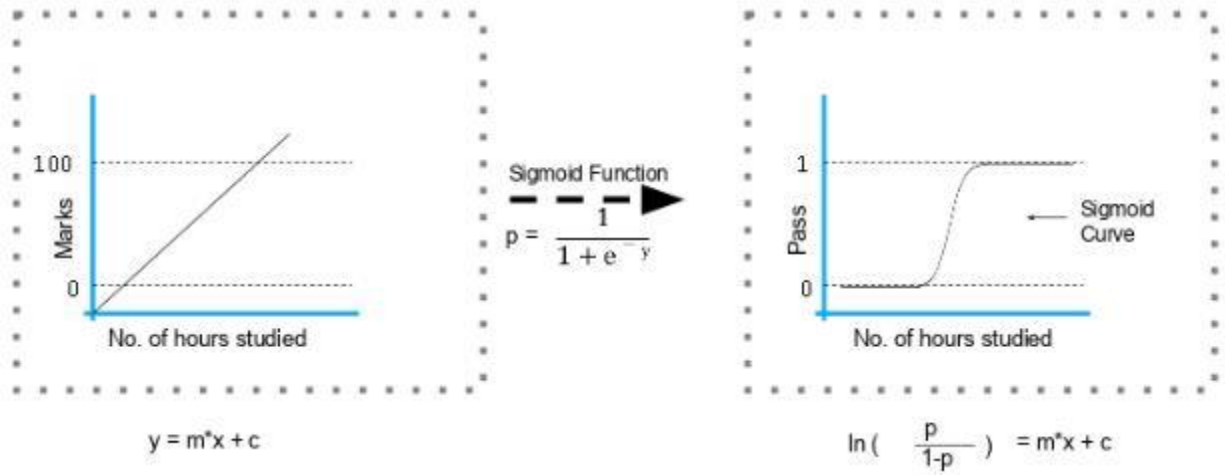
2. How is logistic regression done?

Logistic regression measures the relationship between the dependent variable (our label of what we want to predict) and one or more independent variables (our features) by estimating probability using its underlying logistic function (sigmoid).

The image shown below depicts how logistic regression works:



The formula and graph for the sigmoid function are as shown:



3. Explain the steps in making a decision tree.

Take the entire data set as input

Calculate entropy of the target variable, as well as the predictor attributes

Calculate your information gain of all attributes (we gain information on sorting different objects from each other)

Choose the attribute with the highest information gain as the root node

Repeat the same procedure on every branch until the decision node of each branch is finalized

For example, let's say you want to build a decision tree to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:

It is clear from the decision tree that an offer is accepted if:

Salary is greater than $50,000

The commute is less than an hour

Incentives are offered

A random forest is built up of a number of decision trees. If you split the data into different packages and make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

Randomly select 'k' features from a total of 'm' features where k << m

Among the 'k' features, calculate the node D using the best split point

Split the node into daughter nodes using the best split

Repeat steps two and three until leaf nodes are finalized

Build forest by repeating steps one to four for 'n' times to create 'n' number of trees

5. How can you avoid overfitting your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

Keep the model simple—take fewer variables into account, thereby removing some of the noise in the training data

Use cross-validation techniques, such as k folds cross-validation

Use regularization techniques, such as LASSO, that penalize certain model parameters if they're likely to cause overfitting

6. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Example: height of students

Height (in cm)
164
167.3
170
174.2
178
180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celcius)	Sales
20	2,000
25	2,100
26	2,300
28	2,400
30	2,600
36	3,100

Here, the relationship is visible from the table that temperature and sales are directly proportional to each other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.
Example: data for house price prediction

No. of rooms	Floors	Area (sq ft)	Price
2	0	900	$4000,00
3	2	1,100	$600,000
3.5	5	1,500	$900,000
4	3	2,100	$1,200,000

The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range, minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

7. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection, i.e, filter, and wrapper methods.

Filter Methods

This involves:
Linear discrimination analysis
ANOVA
Chi-Square
The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:
Forward Selection: We test one feature at a time and keep adding them until we get a good fit
Backward Selection: We test all the features and start removing them to see what works better
Recursive Feature Elimination: Recursively looks through all the different features and how they pair together
Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

8. In your choice of language, write a program that prints the numbers ranging from one to 50.

But for multiples of three, print "Fizz" instead of the number, and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz"
The code is shown below:

```
for fizzbuzz in range(51):
    if fizzbuzz % 3 == 0 and fizzbuzz % 5 == 0:
        print("fizzbuzz")
        continue
    elif fizzbuzz % 3 == 0:
        print("fizz")
        continue
    elif fizzbuzz % 5 == 0:
        print("buzz")
        continue
    print(fizzbuzz)
```

Note that the range mentioned is 51, which means zero to 50. However, the range asked in the question is one to 50. Therefore, in the above code, you can include the range as (1,51).
The output of the above code is as shown:

9. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

The following are ways to handle missing data values:

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way; we use the rest of the data to predict the values.

For smaller data sets, we can substitute missing values with the mean or average of the rest of the data using the pandas' data frame in python. There are different ways to do so, such as df.mean(), df.fillna(mean).

10. For the given points, how will you calculate the Euclidean distance in Python?

plot1 = [1,3]

plot2 = [2,5]

The Euclidean distance can be calculated as follows:

euclidean_distance = sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)

Check out the Simplilearn's video on "Data Science Interview Question" curated by industry experts to help you prepare for an interview.

11. What are dimensionality reduction and its benefits?

The Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

12. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

-2	-4	2
-2	1	2
4	2	5

The characteristic equation is as shown:

Expanding determinant:

(-2 – λ) [(1-λ) (5-λ)-2x2] + 4[(-2) x (5-λ) -4x2] + 2[(-2) x 2-4(1-λ)] =0

- λ3 + 4λ2 + 27λ – 90 = 0,

λ3 - 4 λ2 -27 λ + 90 = 0

Here we have an algebraic equation built from the eigenvectors.

By hit and trial:

33 – 4 x 32 - 27 x 3 +90 = 0
Hence, (λ - 3) is a factor:
λ3 - 4 λ2 - 27 λ +90 = (λ – 3) (λ2 – λ – 30)
Eigenvalues are 3,-5,6:
(λ – 3) (λ2 – λ – 30) = (λ – 3) (λ+5) (λ-6),
Calculate eigenvector for λ = 3
For X = 1,
-5 - 4Y + 2Z =0,
-2 - 2Y + 2Z =0
Subtracting the two equations:
3 + 2Y = 0,
Subtracting back into second equation:
Y = -(3/2)
Z = -(1/2)
Similarly, we can calculate the eigenvectors for -5 and 6.

13. How should you maintain a deployed model?

The steps to maintain a deployed model are:

Monitor

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.

Evaluate

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

Compare

The new models are compared to each other to determine which model performs the best.

Rebuild

The best performing model is re-built on the current state of data.

14. What are recommender systems?

A recommender system predicts what a user would rate a specific product based on their preferences. It can be split into two different areas:

Collaborative Filtering

As an example, Last.fm recommends tracks that other users with similar interests play often. This is also commonly seen on Amazon after making a purchase; customers may notice the following message accompanied by product recommendations: "Users who bought this also bought…"

Content-based Filtering

As an example: Pandora uses the properties of a song to recommend music with similar properties. Here, we look at content, instead of looking at who else is listening to music.

15. How do you find RMSE and MSE in a linear regression model?

RMSE and MSE are two of the most common measures of accuracy for a linear regression model.

RMSE indicates the Root Mean Square Error.

```
> rmse
[1] 3.339665e-11
```

MSE indicates the Mean Square Error.

$$MSE = \frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}$$

16. How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run k-means clustering on the data set where 'k' is the number of clusters.

Within the sum of squares (WSS), it is defined as the sum of the squared distance between each member of the cluster and its centroid.

17. What is the significance of p-value?

p-value typically ≤ 0.05

This indicates strong evidence against the null hypothesis; so you reject the null hypothesis.

p-value typically > 0.05

This indicates weak evidence against the null hypothesis, so you accept the null hypothesis.

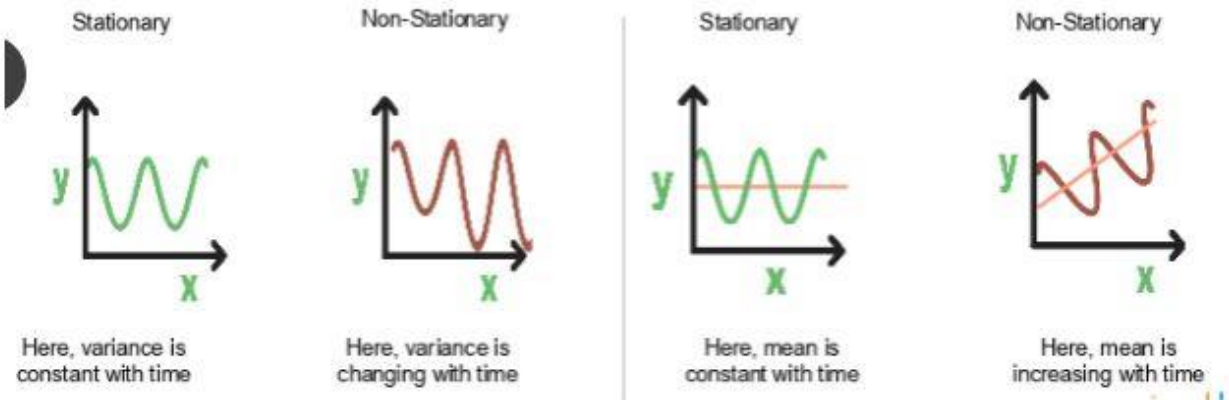
p-value at cutoff 0.05
This is considered to be marginal, meaning it could go either way.

18. How can outlier values be treated?

You can drop outliers only if it is a garbage value.
Example: height of an adult = abc ft. This cannot be true, as the height cannot be a string value. In this case, outliers can be removed.
If the outliers have extreme values, they can be removed. For example, if all the data points are clustered between zero to 10, but one point lies at 100, then we can remove this point.
If you cannot drop outliers, you can try the following:
Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.
Try normalizing the data. This way, the extreme data points are pulled to a similar range.
You can use algorithms that are less affected by outliers; an example would be random forests.

19. How can time-series data be declared as stationery?

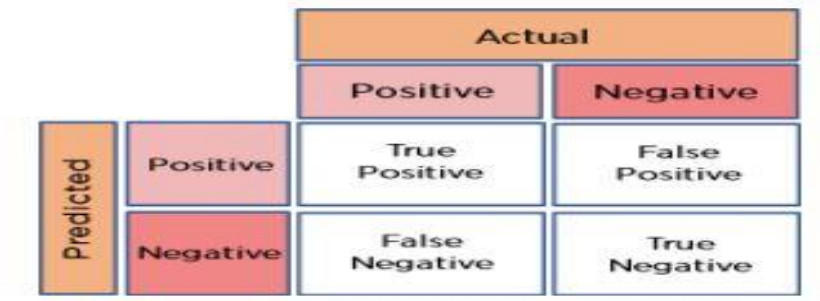
It is stationary when the variance and mean of the series are constant with time.
Here is a visual example:



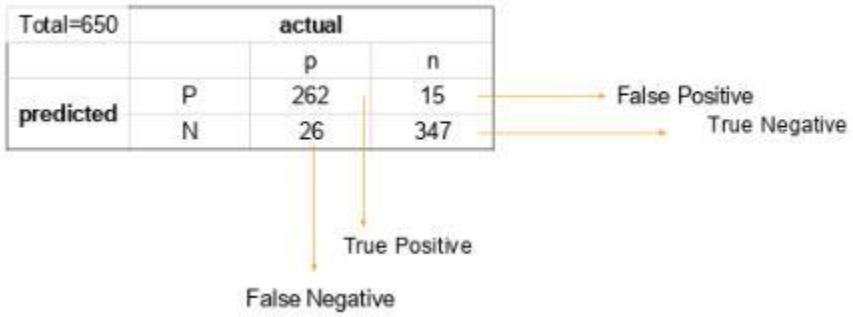
In the first graph, the variance is constant with time. Here, X is the time factor and Y is the variable. The value of Y goes through the same points all the time; in other words, it is stationary.
In the second graph, the waves get bigger, which means it is non-stationary and the variance is changing with time.

Data Science course in Delhi with Placement Guarantee	Data Science Course Dallas	Data Science Course London

20. How can you calculate accuracy using a confusion matrix?
Consider this confusion matrix:



You can see the values for total data, actual values, and predicted values.
The formula for accuracy is:
Accuracy = (True Positive + True Negative) / Total Observations
= (262 + 347) / 650
= 609 / 650
= 0.93
As a result, we get an accuracy of 93 percent.
21. Write the equation and calculate the precision and recall rate.
Consider the same confusion matrix used in the previous question.



Precision = (True positive) / (True Positive + False Positive)
= 262 / 277
= 0.94
Recall Rate = (True Positive) / (Total Positive + False Negative)
= 262 / 288
= 0.90
22. 'People who bought this also bought…' recommendations seen on Amazon are a result of which algorithm?
The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.
The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

For example, a sales page shows that a certain number of people buy a new phone and also buy tempered glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to buy tempered glass as well.

23. Write a basic SQL query that lists all orders with customer information.

Usually, we have order tables and customer tables that contain the following columns:

Order Table
Orderid
customerId
OrderNumber
TotalAmount

Customer Table
Id
FirstName
LastName
City
Country

The SQL query is:

```
SELECT OrderNumber, TotalAmount, FirstName, LastName, City, Country
FROM Order
JOIN Customer
ON Order.CustomerId = Customer.Id
```

24. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection, and can greatly improve a patient's prognosis.

Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

25. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

K-means clustering
Linear regression
K-NN (k-nearest neighbor)
Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance.

26. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable?

[0, 0, 0, 1, 1, 1, 1, 1]

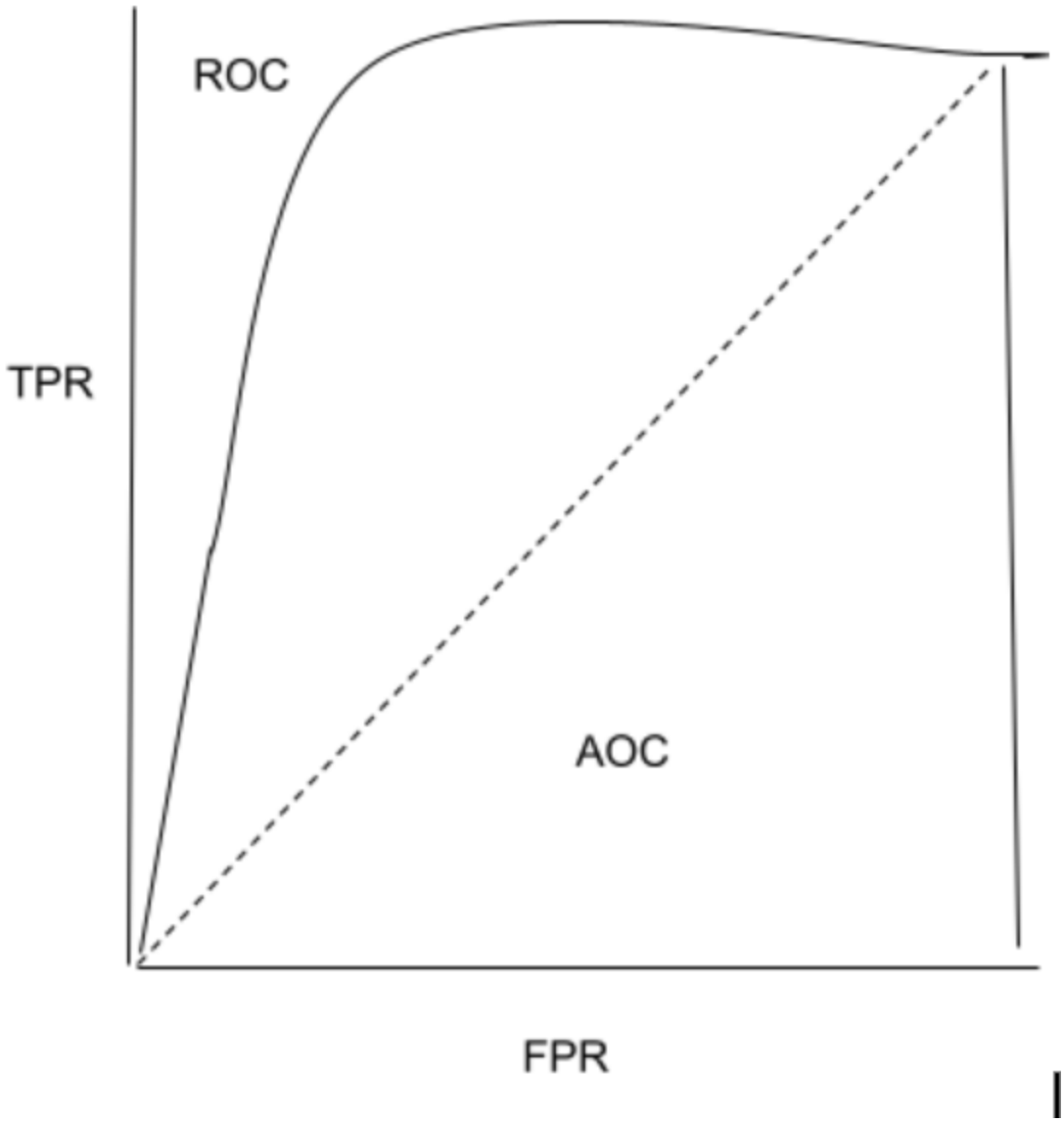
Choose the correct answer.

-(5/8 log(5/8) + 3/8 log(3/8))
5/8 log(5/8) + 3/8 log(3/8)
3/8 log(5/8) + 5/8 log(3/8)
5/8 log(3/8) – 3/8 log(5/8)

The target variable, in this case, is 1.

The formula for calculating the entropy is:

Putting p=5 and n=8, we get
Entropy = A = -(5/8 log(5/8) + 3/8 log(3/8))
27. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case?
Choose the correct option:
Logistic Regression
Linear Regression
K-means clustering
Apriori algorithm
The most appropriate algorithm for this case is A, logistic regression.
28. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?
Choose the correct option:
K-means clustering
Linear regression
Association rules
Decision trees
As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering (answer A) is the most appropriate algorithm for this study.
29. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true?
Choose the right answer:
{banana, apple, grape, orange} must be a frequent itemset
{banana, apple} => {orange} must be a relevant rule
{grape} => {banana, apple} must be a relevant rule
{grape, apple} must be a frequent itemset
The answer is A: {grape, apple} must be a frequent itemset
30. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?
One-way ANOVA
K-means clustering
Association rules
Student's t-test
The answer is A: One-way ANOVA
31. What do you understand about true positive rate and false-positive rate?
The True Positive Rate (TPR) defines the probability that an actual positive will turn out to be positive.
The True Positive Rate (TPR) is calculated by taking the ratio of the [True Positives (TP)] and [True Positive (TP) & False Negatives (FN)].
The formula for the same is stated below -
TPR=TP/TP+FN
The False Positive Rate (FPR) defines the probability that an actual negative result will be shown as a positive one i.e the probability that a model will generate a false alarm.
The False Positive Rate (FPR) is calculated by taking the ratio of the [False Positives (FP)] and [True Positives (TP) & False Positives(FP)].
The formula for the same is stated below -
FPR=FP/TP+FP
32. What is the ROC curve?
The graph between the True Positive Rate on the y-axis and the False Positive Rate on the x-axis is called the ROC curve and is used in binary classification.
The False Positive Rate (FPR) is calculated by taking the ratio between False Positives and the total number of negative samples, and the True Positive Rate (TPR) is calculated by taking the ratio between True Positives and the total number of positive samples.
In order to construct the ROC curve, the TPR and FPR values are plotted on multiple threshold values. The area range under the ROC curve has a range between 0 and 1. A completely random model, which is represented by a straight line, has a 0.5 ROC. The amount of deviation a ROC has from this straight line denotes the efficiency of the model.



The image above denotes a ROC curve example.

Basic Data Science Interview Questions

Let us begin with a few basic data science interview questions!

33. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

34. What are the steps in making a decision tree?

Take the entire data set as input.
Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.
Apply the split to the input data (divide step).
Re-apply steps one and two to the divided data.
Stop when you meet any stopping criteria.
This step is called pruning. Clean up the tree if you went too far doing splits.

35. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

36. What is logistic regression?
Logistic regression is also known as the logit model. It is a technique used to forecast the binary outcome from a linear combination of predictor variables.
37. What are recommender systems?
Recommender systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product.
38. Explain cross-validation.
Cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is to forecast and one wants to estimate how accurately a model will accomplish in practice.
The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) to limit problems like overfitting and gain insight into how the model will generalize to an independent data set.
39. What is collaborative filtering?
Most recommender systems use this filtering process to find patterns and information by collaborating perspectives, numerous data sources, and several agents.
40. Do gradient descent methods always converge to similar points?
They do not, because in some cases, they reach a local minima or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.
41. What is the goal of A/B Testing?
This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.
42. What are the drawbacks of the linear model?
The assumption of linearity of the errors
It can't be used for count outcomes or binary outcomes
There are overfitting problems that it can't solve
43. What is the law of large numbers?
It is a theorem that describes the result of performing the same experiment very frequently. This theorem forms the basis of frequency-style thinking. It states that the sample mean, sample variance, and sample standard deviation converge to what they are trying to estimate.
44. What are the confounding variables?
These are extraneous variables in a statistical model that correlates directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.
45. What is star schema?
It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes, star schemas involve several layers of summarization to recover information faster.
46. How regularly must an algorithm be updated?
You will want to update an algorithm when:
You want the model to evolve as data streams through infrastructure
The underlying data source is changing
There is a case of non-stationarity
47. What are eigenvalue and eigenvector?
Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.
Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix.
48. Why is resampling done?
Resampling is done in any of these cases:
Estimating the accuracy of sample statistics by using subsets of accessible data, or drawing randomly with replacement from a set of data points
Substituting labels on data points when performing significance tests
Validating models by using random subsets (bootstrapping, cross-validation)
49. What is selection bias?
Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.
50. What are the types of biases that can occur during sampling?
Selection bias
Undercoverage bias
Survivorship bias
51. What is survivorship bias?

Survivorship bias is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.

52. How do you work towards a random forest?

The underlying principle of this technique is that several weak learners combine to provide a strong learner. The steps involved are:

Build several decision trees on bootstrapped training samples of data

On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates out of all pp predictors

Rule of thumb: At each split m=p√m=p

Predictions: At the majority rule

This exhaustive list is sure to strengthen your preparation for data science interview questions.

53. What is a bias-variance trade-off?

Bias: Due to an oversimplification of a Machine Learning Algorithm, an error occurs in our model, which is known as Bias. This can lead to an issue of underfitting and might lead to oversimplified assumptions at the model training time to make target functions easier and simpler to understand.

Some of the popular machine learning algorithms which are low on the bias scale are -

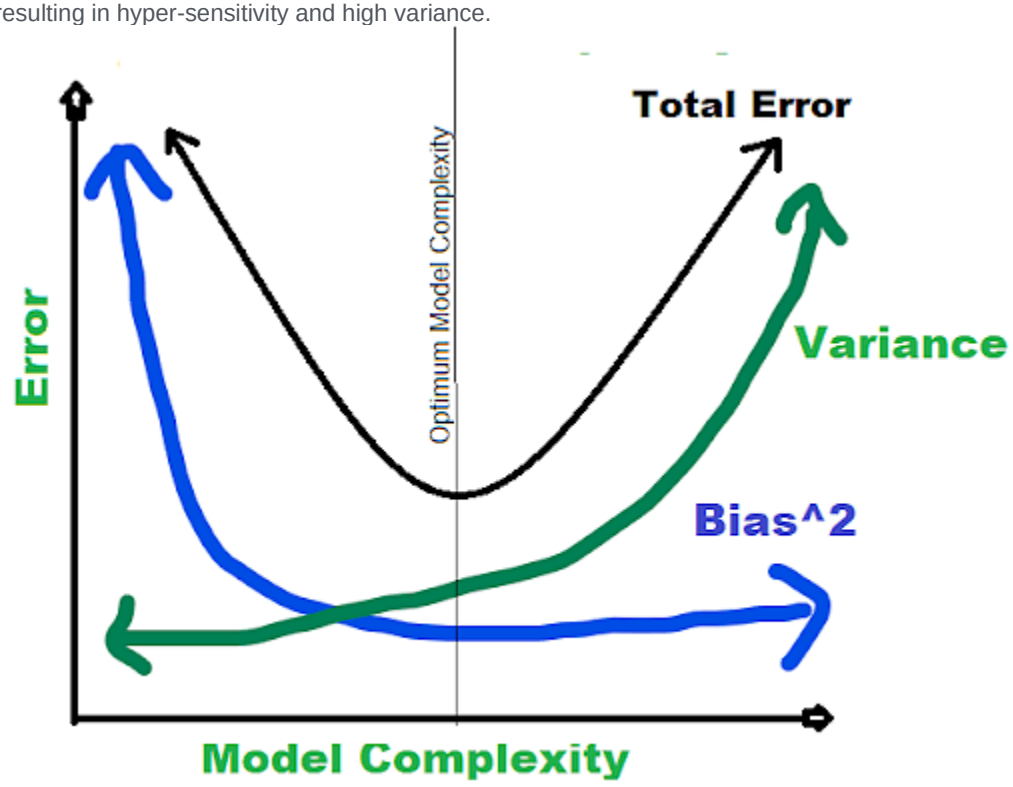
Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees.

Algorithms that are high on the bias scale -

Logistic Regression and Linear Regression.

Variance: Because of a complex machine learning algorithm, a model performs really badly on a test data set as the model learns even noise from the training data set. This error that occurs in the Machine Learning model is called Variance and can generate overfitting and hyper-sensitivity in Machine Learning models.

While trying to get over bias in our model, we try to increase the complexity of the machine learning algorithm. Though it helps in reducing the bias, after a certain point, it generates an overfitting effect on the model hence resulting in hyper-sensitivity and high variance.



Bias-Variance trade-off: To achieve the best performance, the main target of a supervised machine learning algorithm is to have low variance and bias.

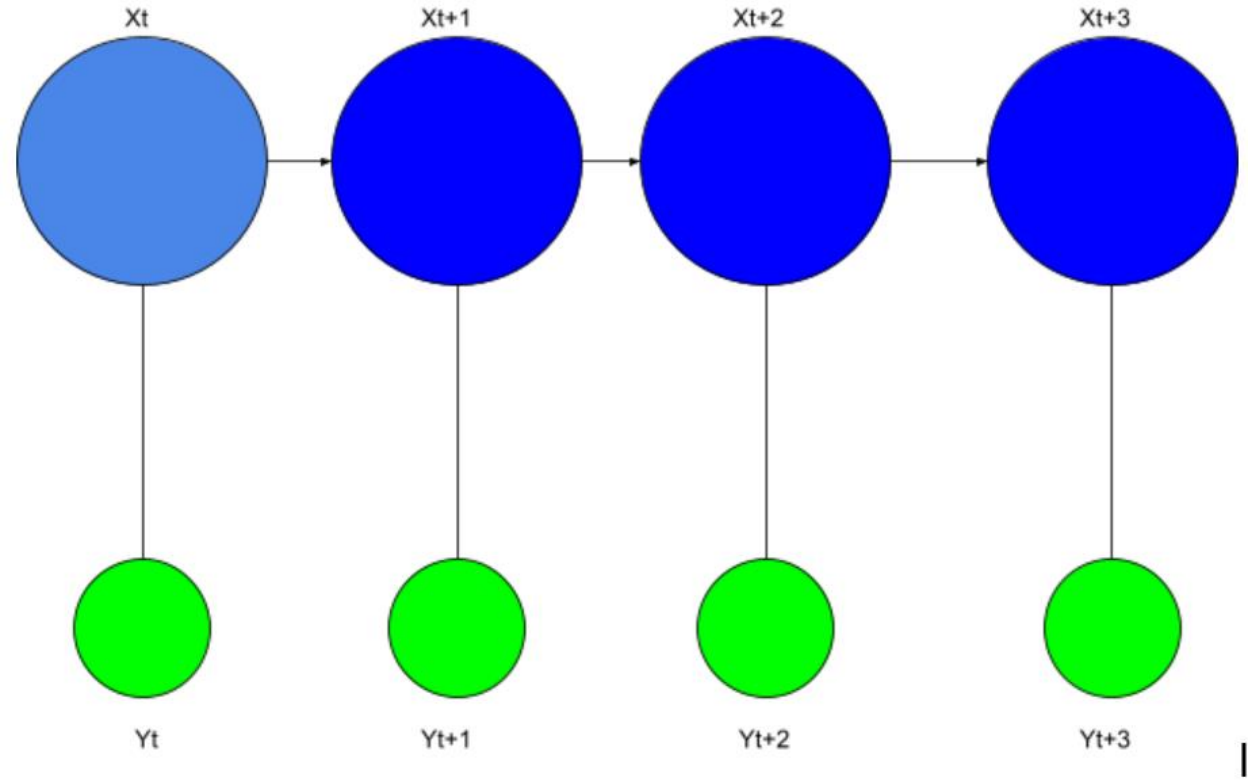
The following things are observed regarding some of the popular machine learning algorithms -

The Support Vector Machine algorithm (SVM) has high variance and low bias. In order to change the trade-off, we can increase the parameter C. The C parameter results in a decrease in the variance and an increase in bias by influencing the margin violations allowed in training datasets.

In contrast to the SVM, the K-Nearest Neighbors (KNN) Machine Learning algorithm has a high variance and low bias. To change the trade-off of this algorithm, we can increase the prediction influencing neighbors by increasing the K value, thus increasing the model bias.

54. Describe Markov chains?

Markov Chains defines that a state's future probability depends only on its current state.
Markov chains belong to the Stochastic process type category.
The below diagram explains a step-by-step model of the Markov Chains whose output depends on their current state.



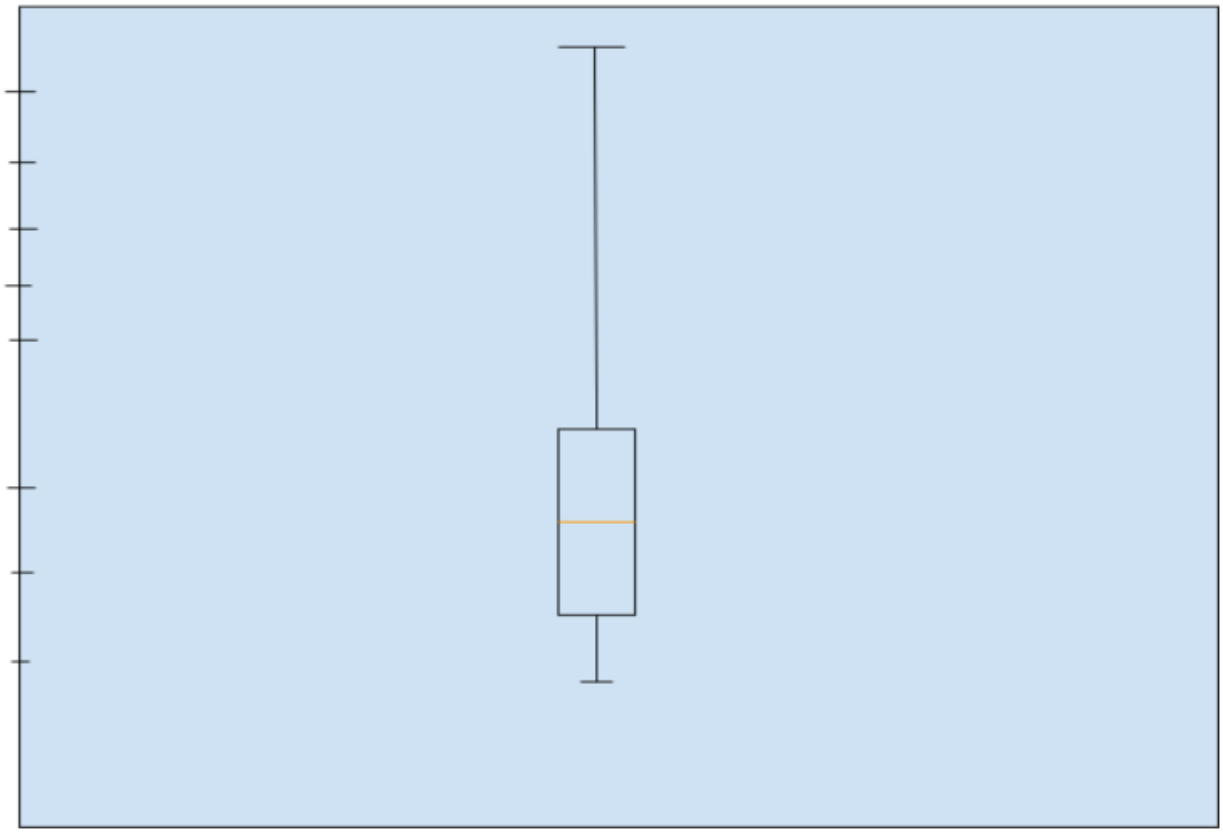
A perfect example of the Markov Chains is the system of word recommendation. In this system, the model recognizes and recommends the next word based on the immediately previous word and not anything before that. The Markov Chains take the previous paragraphs that were similar to training data-sets and generates the recommendations for the current paragraphs accordingly based on the previous word.

55. Why is R used in Data Visualization?

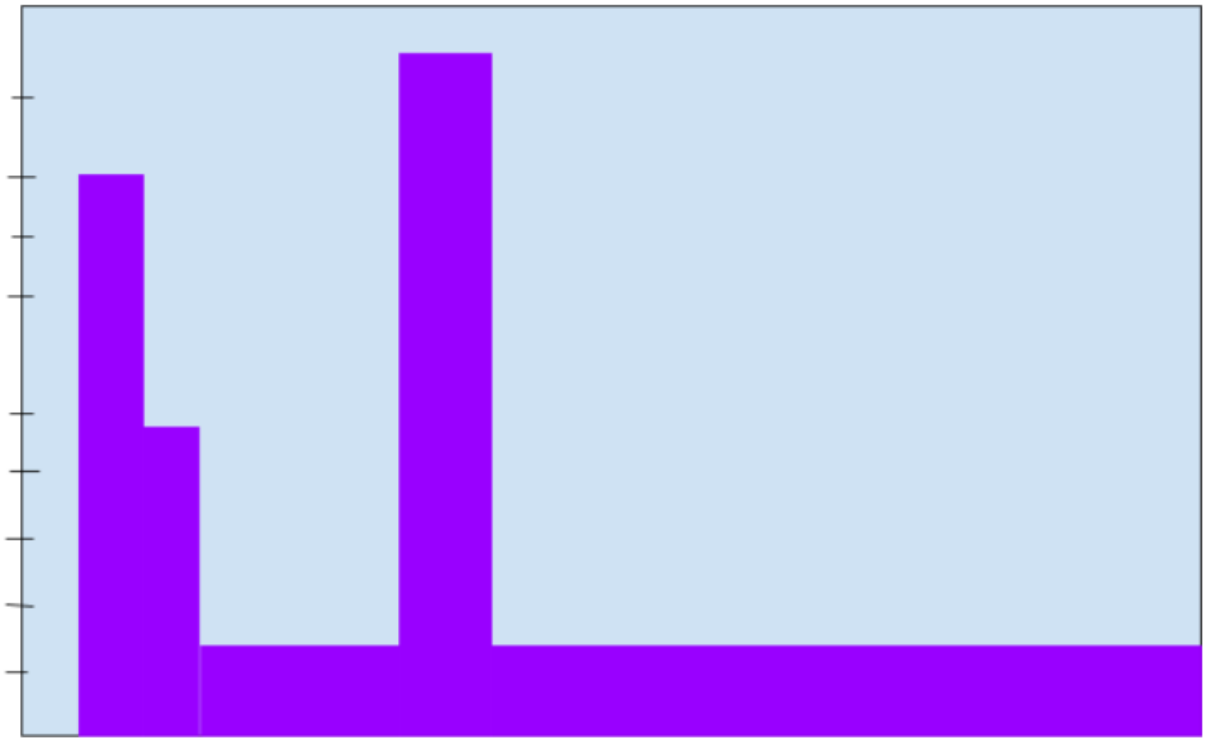
R is widely used in Data Visualizations for the following reasons-
We can create almost any type of graph using R.
R has multiple libraries like lattice, ggplot2, leaflet, etc., and so many inbuilt functions as well.
It is easier to customize graphics in R compared to Python.
R is used in feature engineering and in exploratory data analysis as well.

56. What is the difference between a box plot and a histogram?

The frequency of a certain feature's values is denoted visually by both box plots
and histograms.
Boxplots are more often used in comparing several datasets and compared to histograms, take less space and contain fewer details. Histograms are used to know and understand the probability distribution underlying a dataset.



The diagram above denotes a boxplot of a dataset.



57. What does NLP stand for?

NLP is short for Natural Language Processing. It deals with the study of how computers learn a massive amount of textual data through programming. A few popular examples of NLP are Stemming, Sentimental Analysis, Tokenization, removal of stop words, etc.

58. Difference between an error and a residual error

The difference between a residual error and error are defined below -

Error	Residual Error
The difference between the actual value and the predicted value is called an error. Some of the popular means of calculating data science errors are - Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) Mean Squared Error (MSE)	The difference between the arithmetic mean of a group of values and the observed group of values is called a residual error.
An error is generally unobservable.	A residual error can be represented using a graph.

A residual error is used to show how the sample population data and the observed data differ from each other.	An error is how actual population data and observed data differ from each other.

59. Difference between Normalisation and Standardization

Standardization	Normalization
The technique of converting data in such a way that it is normally distributed and has a standard deviation of 1 and a mean of 0.	The technique of converting all data values to lie between 1 and 0 is known as Normalization. This is also known as min-max scaling.
Standardization takes care that the standard normal distribution is followed by the data.	The data returning into the 0 to 1 range is taken care of by Normalization.
Normalization formula - X' = (X - Xmin) / (Xmax - Xmin) Here, Xmin - feature's minimum value, Xmax - feature's maximum value.	Standardization formula - X' = (X - μ) / σ

60. Difference between Point Estimates and Confidence Interval

Confidence Interval: A range of values likely containing the population parameter is given by the confidence interval. Further, it even tells us how likely that particular interval can contain the population parameter. The Confidence Coefficient (or Confidence level) is denoted by 1-alpha, which gives the probability or likeness. The level of significance is given by alpha.
Point Estimates: An estimate of the population parameter is given by a particular value called the point estimate. Some popular methods used to derive Population Parameters' Point estimators are - Maximum Likelihood estimator and the Method of Moments.
To conclude, the bias and variance are inversely proportional to each other, i.e., an increase in bias results in a decrease in the variance, and an increase in variance results in a decrease in bias.

Machine Learning

1) What is Machine learning?
Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience. For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.
2) Mention the difference between Data Mining and Machine learning?
Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While, data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this process machine, learning algorithms are used.

3) What is 'Overfitting' in Machine learning?
In machine learning, when a statistical model describes random error or noise instead of underlying relationship 'overfitting' occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.
4) Why overfitting happens?
The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.
5) How can you avoid overfitting ?
By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.
In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to "test" the model in the training phase.
6) What is inductive machine learning?
The inductive machine learning involves the process of learning by examples, where a system, from a set of observed instances tries to induce a general rule.
7) What are the five popular algorithms of Machine Learning?
a) Decision Trees
b) Neural Networks (back propagation)
c) Probabilistic networks
d) Nearest Neighbor
e) Support vector machines
8) What are the different Algorithm techniques in Machine Learning?
The different types of techniques in Machine Learning are
a) Supervised Learning
b) Unsupervised Learning
c) Semi-supervised Learning
d) Reinforcement Learning
e) Transduction
f) Learning to Learn
9) What are the three stages to build the hypotheses or model in machine learning?
a) Model building
b) Model testing
c) Applying the model
10) What is the standard approach to supervised learning?
The standard approach to supervised learning is to split the set of example into the training set and the test.
11) What is 'Training set' and 'Test set'?
In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.
12) List down various approaches for machine learning?
The different approaches in Machine Learning are
a) Concept Vs Classification Learning
b) Symbolic Vs Statistical Learning
c) Inductive Vs Analytical Learning
13) What is not Machine Learning?
a) Artificial Intelligence
b) Rule based inference
14) Explain what is the function of 'Unsupervised Learning'?
a) Find clusters of the data
b) Find low-dimensional representations of the data
c) Find interesting directions in data

d) Interesting coordinates and correlations
e) Find novel observations/ database cleaning

15) Explain what is the function of 'Supervised Learning'?

a) Classifications
b) Speech recognition
c) Regression
d) Predict time series
e) Annotate strings

16) What is algorithm independent machine learning?

Machine learning in where mathematical foundations is independent of any particular classifier or learning algorithm is referred as algorithm independent machine learning?

17) What is the difference between artificial learning and machine learning?

Designing and developing algorithms according to the behaviours based on empirical data are known as Machine Learning. While artificial intelligence in addition to machine learning, it also covers other aspects like knowledge representation, natural language processing, planning, robotics etc.

18) What is classifier in machine learning?

A classifier in a Machine Learning is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.

19) What are the advantages of Naive Bayes?

In Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. The main advantage is that it can't learn interactions between features.

20) In what areas Pattern Recognition is used?

Pattern Recognition can be used in
a) Computer Vision
b) Speech Recognition
c) Data Mining
d) Statistics
e) Informal Retrieval
f) Bio-Informatics

21) What is Genetic Programming?

Genetic programming is one of the two techniques used in machine learning. The model is based on the testing and selecting the best choice among a set of results.

22) What is Inductive Logic Programming in Machine Learning?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical programming representing background knowledge and examples.

23) What is Model Selection in Machine Learning?

The process of selecting models among different mathematical models, which are used to describe the same data set is known as Model Selection. Model selection is applied to the fields of statistics, machine learning and data mining.

24) What are the two methods used for the calibration in Supervised Learning?

The two methods used for predicting good probabilities in Supervised Learning are
a) Platt Calibration
b) Isotonic Regression
These methods are designed for binary classification, and it is not trivial.

25) Which method is frequently used to prevent overfitting?

When there is sufficient data 'Isotonic Regression' is used to prevent an overfitting issue.

26) What is the difference between heuristic for rule learning and heuristics for decision trees?

The difference is that the heuristics for decision trees evaluate the average quality of a number of disjointed sets while rule learners only evaluate the quality of the set of instances that is covered with the candidate rule.

27) What is Perceptron in Machine Learning?

In Machine Learning, Perceptron is an algorithm for supervised classification of the input into one of

several possible non-binary outputs.

28) Explain the two components of Bayesian logic program?

Bayesian logic program consists of two components. The first component is a logical one ; it consists of a set of Bayesian Clauses, which captures the qualitative structure of the domain. The second component is a quantitative one, it encodes the quantitative information about the domain.

29) What are Bayesian Networks (BN) ?

Bayesian Network is used to represent the graphical model for probability relationship among a set of variables .

30) Why instance based learning algorithm sometimes referred as Lazy learning algorithm?

Instance based learning algorithm is also referred as Lazy learning algorithm as they delay the induction or generalization process until classification is performed.

31) What are the two classification methods that SVM (Support Vector Machine) can handle?

a) Combining binary classifiers

b) Modifying binary to incorporate multiclass learning

32) What is ensemble learning?

To solve a particular computational program, multiple models such as classifiers or experts are strategically generated and combined. This process is known as ensemble learning.

33) Why ensemble learning is used?

Ensemble learning is used to improve the classification, prediction, function approximation etc of a model.

34) When to use ensemble learning?

Ensemble learning is used when you build component classifiers that are more accurate and independent from each other.

35) What are the two paradigms of ensemble methods?

The two paradigms of ensemble methods are

a) Sequential ensemble methods

b) Parallel ensemble methods

36) What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

37) What is bias-variance decomposition of classification error in ensemble method?

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

38) What is an Incremental Learning algorithm in ensemble?

Incremental learning method is the ability of an algorithm to learn from new data that may be available after classifier has already been generated from already available dataset.

39) What is PCA, KPCA and ICA used for?

PCA (Principal Components Analysis), KPCA (Kernel based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

40) What is dimension reduction in Machine Learning?

In Machine Learning and statistics, dimension reduction is the process of reducing the number of random variables under considerations and can be divided into feature selection and feature extraction

41) What are support vector machines?

Support vector machines are supervised learning algorithms used for classification and regression analysis.

42) What are the components of relational evaluation techniques?

The important components of relational evaluation techniques are

a) Data Acquisition
b) Ground Truth Acquisition
c) Cross Validation Technique
d) Query Type
e) Scoring Metric
f) Significance Test

43) What are the different methods for Sequential Supervised Learning?

The different methods to solve Sequential Supervised Learning problems are

a) Sliding-window methods
b) Recurrent sliding windows
c) Hidden Markow models
d) Maximum entropy Markow models
e) Conditional random fields
f) Graph transformer networks

44) What are the areas in robotics and information processing where sequential prediction problem arises?

The areas in robotics and information processing where sequential prediction problem arises are

a) Imitation Learning
b) Structured prediction
c) Model based reinforcement learning

45) What is batch statistical learning?

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

46) What is PAC Learning?

PAC (Probably Approximately Correct) learning is a learning framework that has been introduced to analyze learning algorithms and their statistical efficiency.

47) What are the different categories you can categorized the sequence learning process?

a) Sequence prediction
b) Sequence generation
c) Sequence recognition
d) Sequential decision

48) What is sequence learning?

Sequence learning is a method of teaching and learning in a logical manner.

49) What are two techniques of Machine Learning ?

The two techniques of Machine Learning are

a) Genetic Programming
b) Inductive Learning

50) Give a popular application of machine learning that you see on day to day basis?

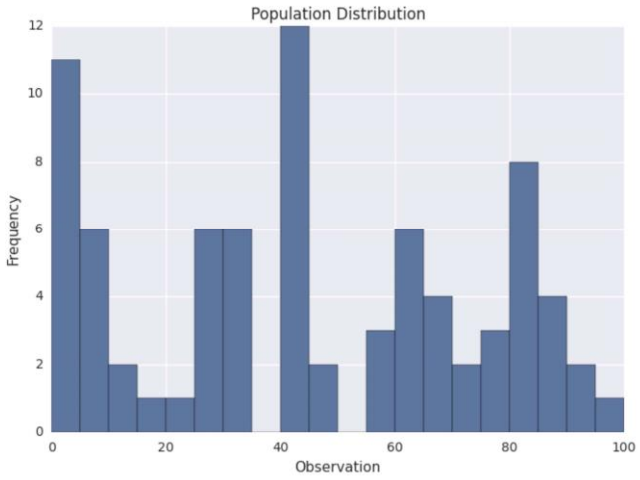
The recommendation engine implemented by major ecommerce websites uses Machine Learning

Data Science interview questions

Statistics

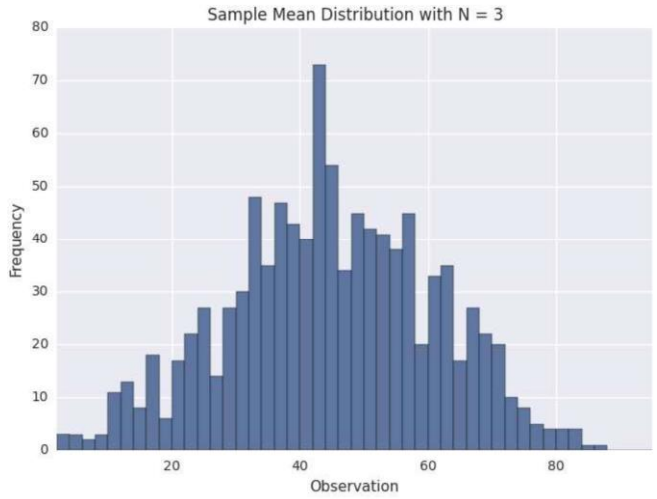
Q1. What is the Central Limit Theorem and why is it important?

Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impractical, bordering on impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly. Formally, it states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling), the mean tending to the mean
of the population and variance equal to the variance of the population divided by the size of the sampling.
What's especially important is that this will be true regardless of the distribution of the original population.

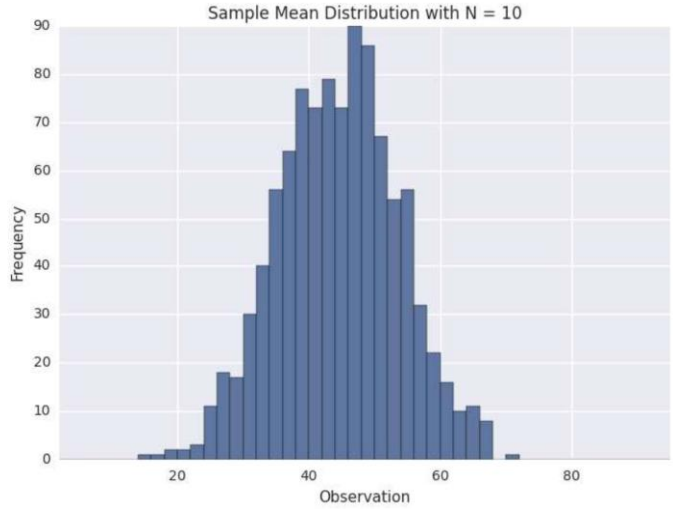


EX:

As we can see, the distribution is pretty ugly. It certainly isn't normal, uniform, or any other commonly known distribution. In order to sample from the above distribution, we need to define a sample size, referred to as N. This is the number of observations that we will sample at a time. Suppose that we choose
N to be 3. This means that we will sample in groups of 3. So for the above population, we might sample groups such as [5, 20, 41], [60, 17, 82], [8, 13, 61], and so on.
Suppose that we gather 1,000 samples of 3 from the above population. For each sample, we can compute
its average. If we do that, we will have 1,000 averages. This set of 1,000 averages is called a sampling distribution, and according to Central Limit Theorem, the sampling distribution will approach a normal distribution as the sample size N used to produce it increases. Here is what our sample distribution looks like for N = 3.



As we can see, it certainly looks uni-modal, though not necessarily normal. If we repeat the same process
with a larger sample size, we should see the sampling distribution start to become more normal. Let's repeat the same process again with N = 10. Here is the sampling distribution for that sample size.



Q2. What is sampling? How many sampling methods do you know?

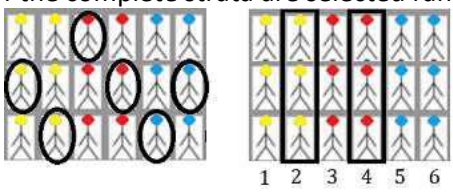
Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined. It enables data
scientists, predictive modelers and other data analysts to work with a small, manageable amount of data about a statistical population to build and run analytical models more quickly, while still producing accurate findings.

Sampling can be particularly useful with data sets that are too large to efficiently analyze in full – for example, in big data analytics applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.

An important consideration, though, is the size of the required data sample and the possibility of introducing a sampling error. In some cases, a small sample can reveal the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing

the data as a whole, even though the increased size of the sample may impede ease of manipulation and
interpretation.
There are many different methods for drawing samples from data; the ideal one depends on the data set
and situation. Sampling can be based on probability, an approach that uses random numbers that correspond to points in the data set to ensure that there is no correlation between points chosen for the sample. Further variations in probability sampling include:

- Simple random sampling: Software is used to randomly select subjects from the whole population.
- Stratified sampling: Subsets of the data sets or population are created based on a common factor, and samples are randomly collected from each subgroup. A sample is drawn from each strata (using a random sampling method like simple random sampling or systematic sampling).
  - EX: In the image below, let's say you need a sample size of 6. Two members from each group (yellow, red, and blue) are selected randomly. Make sure to sample proportionally: In this simple example, 1/3 of each group (2/6 yellow, 2/6 red and 2/6 blue) has been sampled. If you have one group that's a different size, make sure to adjust your proportions. For example, if you had 9 yellow, 3 red and 3 blue, a 5-item sample would consist of 3/9 yellow (i.e. one third), 1/3 red and 1/3 blue.
- Cluster sampling: The larger data set is divided into subsets (clusters) based on a defined factor, then a random sampling of clusters is analyzed. The sampling unit is the whole cluster; Instead of sampling individuals from within each group, a researcher will study whole clusters.
  - EX: In the image below, the strata are natural groupings by head color (yellow, red, blue). A sample size of 6 is needed, so two of the complete strata are selected randomly (in this example, groups 2 and 4 are chosen).



Stratified Sampling **Cluster Sampling**

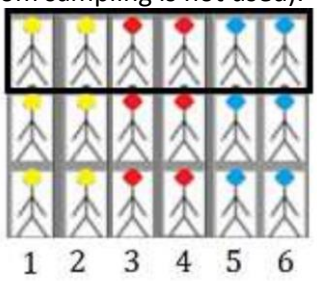
- Multistage sampling: A more complicated form of cluster sampling, this method also involves dividing the larger population into a number of clusters. Second-stage clusters are then broken out based on a secondary factor, and those clusters are then sampled and analyzed. This staging could continue as multiple subsets are identified, clustered and analyzed.
- Systematic sampling: A sample is created by setting an interval at which to extract data from the larger population – for example, selecting every 10th row in a spreadsheet of 200 items to create a sample size of 20 rows to analyze.

Sampling can also be based on non-probability, an approach in which a data sample is determined and extracted based on the judgment of the analyst. As inclusion is determined by the analyst, it can be more
difficult to extrapolate whether the sample accurately represents the larger population than when probability sampling is used.
Non-probability data sampling methods include:

- Convenience sampling: Data is collected from an easily accessible and available group.
- Consecutive sampling: Data is collected from every subject that meets the criteria until the predetermined sample size is met.
- Purposive or judgmental sampling: The researcher selects the data to sample based on predefined

criteria.

• Quota sampling: The researcher ensures equal representation within the sample for all subgroups in the data set or population (random sampling is not used).



Quota Sampling

Once generated, a sample can be used for predictive analytics. For example, a retail business might use data sampling to uncover patterns about customer behavior and predictive modeling to create more effective sales strategies.

Q3. What is the difference between type I vs type II error?

Is H_a true? No, H_0 is True (H_a is Negative: TN); Yes, H_0 is False (H_a is Positive: TP).

A type I error occurs when the null hypothesis is true but is rejected. A type II error occurs when the null hypothesis is false but erroneously fails to be rejected.

TN FP (I error)
FN (II error) TP

	No reject H_0	Reject H_0
H_0 is True	TN	FP (I error)
H_0 is False	FN (II error)	TP

Q4. What is linear regression? What do the terms p-value, coefficient, and rsquared value mean? What is the significance of each of these components?

Imagine you want to predict the price of a house. That will depend on some factors, called independent variables, such as location, size, year of construction… if we assume there is a linear relationship between
these variables and the price (our dependent variable), then our price is predicted by the following function:

= □ + □ □

The p-value in the table is the minimum □ (the significance level) at which the coefficient is relevant. The lower the p-value, the more important is the variable in predicting the price. Usually we set a 5% level, so
that we have a 95% confidentiality that our variable is relevant.

The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable

in isolation from the others.
R squared (R_2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Q5. What are the assumptions required for linear regression?

There are four major assumptions:

There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data,
The errors or residuals ($y_i - \hat{y}_i$) of the data are normally distributed and independent from each other,
There is minimal multicollinearity between explanatory variables, and
Homoscedasticity. This means the variance around the regression line is the same for all values

• There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data,
• The errors or residuals (□□ – □□□) of the data are normally distributed and independent from each other,
• There is minimal multicollinearity between explanatory variables, and
• Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

Q6. What is a statistical interaction?

Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output
variable) differs among levels of another factor. When two or more independent variables are involved in
a research design, there is more to consider than simply the "main effect" of each of the independent variables (also termed "factors"). That is, the effect of one independent variable on the dependent variable of interest may not be the same at all levels of the other independent variable. Another way to put this is that the effect of one independent variable may depend on the level of the other independent
variable. In order to find an interaction, you must have a factorial design, in which the two (or more) independent variables are "crossed" with one another so that there are observations at every combination of levels of the two independent variables. EX: stress level and practice to memorize words:
together they may have a lower performance.

Q7. What is selection bias?

Selection (or 'sampling') bias occurs when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data is systematically (i.e., non-randomly) excluded from analysis.

Q8. What is an example of a data set with a non-Gaussian distribution?

The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has
a solid grounding in statistics, they can be utilized where appropriate.
Binomial: multiple toss of a coin Bin(n,p): the binomial distribution consists of the probabilities of each of

the possible numbers of successes on n trials for independent events that each have a probability of p of occurring.
Bernoulli: Bin(1,p) = Be(p)
Poisson: Pois(□)

Data Science

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing
for years? The answer lies in the difference between explaining and predicting: statisticians work a posteriori, explaining the results and designing a plan; data scientists use historical data to make predictions.
The differences between supervised and unsupervised learning are:

Supervised	Unsupervised
Input data is labelled	Input data is unlabeled
Split in training/validation/test	No split
Used for prediction	Used for analysis
Classification and Regression	Clustering, dimension reduction, and density estimation

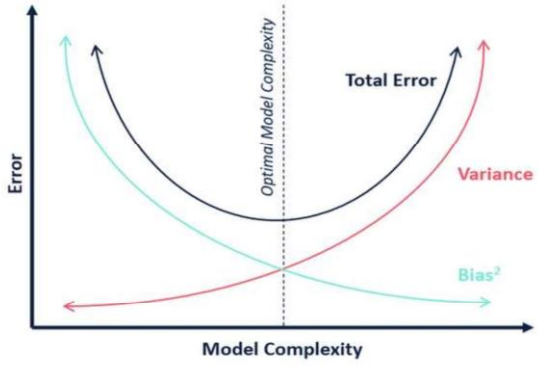
Q2. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides what has to be studied. It is associated with research where the selection of participants is not random. Therefore, some conclusions of the study may not be accurate.
The types of selection bias include:
• Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
• Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
• Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
• Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

Q3. What is bias-variance trade-off?

Bias: Bias is an error introduced in the model due to the oversimplification of the algorithm used (does not fit the data properly). It can lead to under-fitting.
Low bias machine learning algorithms — Decision Trees, k-NN and SVM
High bias machine learning algorithms — Linear Regression, Logistic Regression
Variance: Variance is error introduced in the model due to a too complex algorithm, it performs very well
in the training set but poorly in the test set. It can lead to high sensitivity and overfitting.
Possible high variance – polynomial regression

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model
more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.
1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.
3. The decision tree has low bias and high variance, you can decrease the depth of the tree or use fewer attributes.
4. The linear regression has low variance and high bias, you can increase the number of features or use another regression that better fits the data.
There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

Q4. What is a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the binary classifier.

	Predict +	Predict -
Actual +	TP	FN (II error)
Actual -	FP (I error)	TN

A data set used for performance evaluation is called a test data set. It should contain the correct labels and predicted labels. The predicted labels will exactly the same if the performance of a binary classifier is
perfect. The predicted labels usually match with part of the observed labels in real-world scenarios.
A binary classifier predicts all data instances of a test data set as either positive or negative. This produces
four outcomes: TP, FP, TN, FN. Basic measures derived from the confusion matrix:

1. $Error\ Rate\ = \frac{FP+FN}{P+N}$
2. $Accuracy\ = \frac{TP+T}{P+N}$
3. $Sensitivity\ (Recall\ or\ True\ positive\ rate) = \frac{TP}{TP+FN} = \frac{TP}{P}$
4. $Specificity\ (True\ negative\ rate) = \frac{TN}{TN+FP} = \frac{TN}{N}$
5. $Precision\ (Positive\ predicted\ value) = \frac{TP}{TP+FP}$
6. $F-Score\ (Harmonic\ mean\ of\ precision\ and\ recall) = \frac{2\,TP}{(2\,TP+FP+FN)}$

Q5. What is the difference between “long” and “wide” format data?

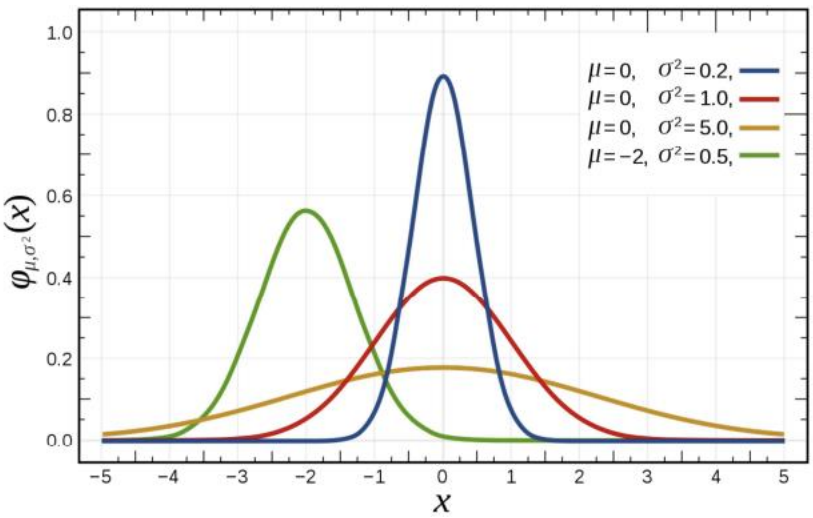
In the wide-format, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the long-format, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups (variables).

X	Y1	Y2	Y3
10	2	3	4
15	0	4	6
20	1	4	5

VarName	X	Value
Y1	10	2
Y2	10	3
Y3	10	4
Y1	15	0
Y2	15	4
Y3	15	6
Y1	20	1
Y2	20	4
Y3	20	5

Q6. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left
or right and reaches normal distribution in the form of a bell-shaped curve.



The random variables are distributed in the form of a symmetrical, bell-shaped curve. Properties of Normal Distribution are as follows:

1. Unimodal (Only one mode)
2. Symmetrical (left and right halves are mirror images)
3. Bell-shaped (maximum height (mode) at the mean)
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

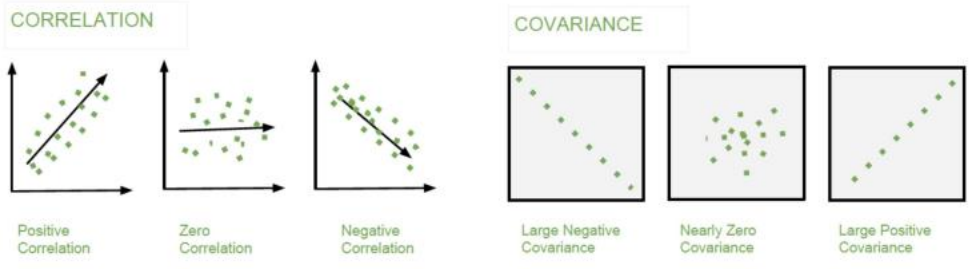
Q7. What is correlation and covariance in statistics?

Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Given two random variables, it is the covariance between both divided by the product of the two
standard deviations of the single variables, hence always between -1 and 1.

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)} \in [-1,1]$$

Covariance is a measure that indicates the extent to which two random variables change in cycle. It explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$



Q8. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments
and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.
A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population
parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by $1 - \alpha$, where α is the level of significance.

Q9. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.
The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search
ads. An example of this could be identifying the click-through rate for a banner ad.

Q10. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is the minimum significance level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

Q11. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

- $Probability\ of\ not\ seeing\ any\ shooting\ star\ in\ 15\ minutes\ is\ = 1 - P(Seeing\ one\ shooting\ star) = 1 - 0.2 = 0.8$
- $Probability\ of\ not\ seeing\ any\ shooting\ star\ in\ the\ period\ of\ one\ hour = (0.8)^4 = 0.4096$

- $Probability\ of\ seeing\ at\ least\ one\ shooting\ star\ in\ the\ one\ hour\ = 1 - P(Not\ seeing\ any\ star) = 1 - 0.4096 = 0.5904$

Q12. How can you generate a random number between 1 – 7 with only a die?

Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes. To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35
outcomes and exclude the other one. A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice. All the remaining combinations from (1,1) till (6,5) can be divided into
7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

Q13. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

$$P(Having\ two\ girls\ given\ one\ girl) = \frac{1}{2}$$

Q14. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

$$Probability\ of\ selecting\ fair\ coin\ = \frac{999}{1000}\ =\ 0.999$$

$$Probability\ of\ selecting\ unfair\ coin\ = \frac{1}{1000} =\ 0.001$$

$$\begin{aligned} &Selecting\ 10\ heads\ in\ a\ row \\ &= Selecting\ fair\ coin\ *\ Getting\ 10\ heads\ +\ Selecting\ unfair\ coin \\ &= P(A) + P(B) \end{aligned}$$

$$P(A) = 0.999 * \left(\frac{1}{2}\right)^{10} = 0.999 * \left(\frac{1}{1024}\right) = 0.000976$$

$$P(B) = 0.001 * 1 = 0.001$$

$$\frac{P(A)}{P(A) + P(B)} = \frac{0.000976}{0.000976 + 0.001} = 0.4939$$

$$\frac{P(B)}{P(A) + P(B)} = \frac{0.001}{0.001976} = 0.5061$$

$$Probability\ of\ selecting\ another\ head\ = \frac{P(A)}{P(A) + P(B)} * 0.5 + \frac{P(B)}{P(A) + P(B)} * 1 =$$
$$= 0.4939 * 0.5 + 0.5061 = 0.7531$$

Q15. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

$$Sensitivity\ = \frac{TP}{TP + FN}$$

Q16. Why is Re-sampling done?

https://machinelearningmastery.com/statistical-sampling-and-resampling/

• Sampling is an active process of gathering observations with the intent of estimating a population variable.

• Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.

Once we have a data sample, it can be used to estimate the population parameter. The problem is that we only have a single estimate of the population parameter, with little idea of the variability or uncertainty

in the estimate. One way to address this is by estimating the population parameter multiple times from our data sample. This is called resampling. Statistical resampling methods are procedures that describe how to economically use available data to estimate a population parameter. The result can be both a more accurate estimate of the parameter (such as taking the mean of the estimates) and a quantification

of the uncertainty of the estimate (such as adding a confidence interval).

Resampling methods are very easy to use, requiring little mathematical knowledge. A downside of the

methods is that they can be computationally very expensive, requiring tens, hundreds, or even thousands
of resamples in order to develop a robust estimate of the population parameter.
The key idea is to resample form the original data — either directly or via a fitted model — to create replicate datasets, from which the variability of the quantiles of interest can be assessed without longwinded
and error-prone analytical calculation. Because this approach involves repeating the original data analysis procedure with many replicate sets of data, these are sometimes called computer-intensive methods. Each new subsample from the original data sample is used to estimate the population parameter. The sample of estimated population parameters can then be considered with statistical tools in order to quantify the expected value and variance, providing measures of the uncertainty of the estimate. Statistical sampling methods can be used in the selection of a subsample from the original sample.
A key difference is that process must be repeated multiple times. The problem with this is that there will be some relationship between the samples as observations that will be shared across multiple subsamples. This means that the subsamples and the estimated population parameters are not strictly identical and independently distributed. This has implications for statistical tests performed on the sample
of estimated population parameters downstream, i.e. paired statistical tests may be required.
Two commonly used resampling methods that you may encounter are k-fold cross-validation and the bootstrap.
• Bootstrap. Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.
• k-fold Cross-Validation. A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set.
The k-fold cross-validation method specifically lends itself to use in the evaluation of predictive models that are repeatedly trained on one subset of the data and evaluated on a second held-out subset of the data.
Resampling is done in any of these cases:
• Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
• Substituting labels on data points when performing significance tests
• Validating models by using random subsets (bootstrapping, cross-validation)

Q17. What are the differences between over-fitting and under-fitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data,
so as to be able to make reliable predictions on general untrained data.
In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.
Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying
trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Q18. How to combat Overfitting and Underfitting?

To combat overfitting:
1. Add noise
2. Feature selection
3. Increase training set
4. L2 (ridge) or L1 (lasso) regularization; L1 drops weights, L2 no
5. Use cross-validation techniques, such as k folds cross-validation
6. Boosting and bagging
7. Dropout technique
8. Perform early stopping
9. Remove inner layers

To combat underfitting:
1. Add features
2. Increase time of training

Q19. What is regularization? Why is it useful?

Regularization is the process of adding tuning parameter (penalty term) to a model to induce smoothness
in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight
vector. This constant is often the L1 (Lasso - |□|) or L2 (Ridge - □□). The model predictions should then minimize the loss function calculated on the regularized training set.

Q20. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This
theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

Q21. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.
If you are researching whether a lack of exercise leads to weight gain:
lack of exercise = independent variable
weight gain = dependent variable
A confounding variable here would be any other variable that affects both of these variables, such as the age of the subject.

Q22. What Are the Types of Biases That Can Occur During Sampling?

a. Selection bias
b. Under coverage bias
c. Survivorship bias

Q23. What is Survivorship Bias?

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means. For example, during a recession you look just at the survived businesses, noting
that they are performing poorly. However, they perform better than the rest, which is failed, thus being removed from the time series.

Q24. What is Selection Bias? What is under coverage bias?

https://stattrek.com/survey-research/survey-bias.aspx

Selection bias occurs when the sample obtained is not representative of the population intended to be analyzed. For instance, you select only Asians to perform a study on the world population height.

Under coverage bias occurs when some members of the population are inadequately represented in the sample. A classic example of under coverage is the Literary Digest voter survey, which predicted that Alfred

Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from

under coverage of low-income voters, who tended to be Democrats.

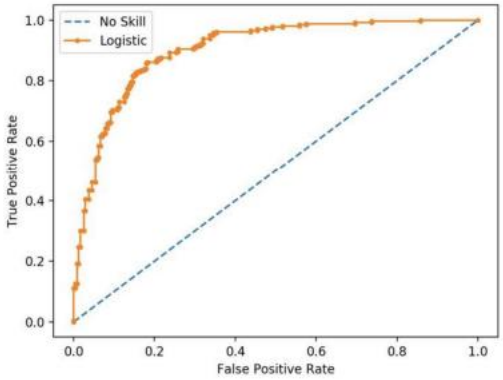
How did this happen? The survey relied on a convenience sample, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent. Under coverage is often a problem with convenience samples.

Q25. Explain how a ROC curve works?

The ROC curve is a graphical representation of the contrast between true positive rates and false positive

rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity (true positive rate) and false positive rate.

- $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$
- $TNR = \frac{TN}{TN+FP} = \frac{TN}{N}$
- $FPR = \frac{FP}{TN+FP}$
- $FNR = \frac{FN}{FN+T}$



Q26. What is TF/IDF vectorization?

TF-IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to

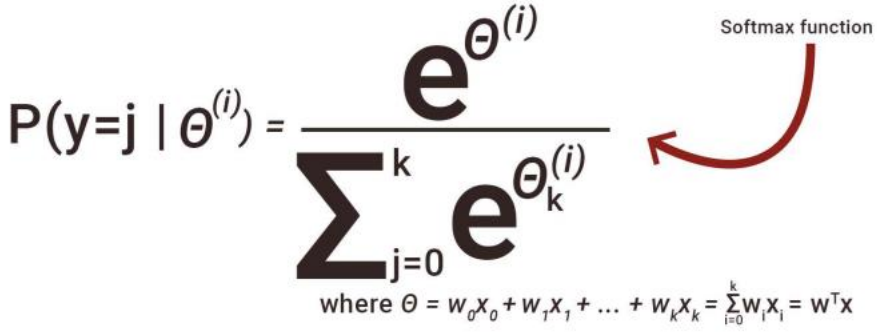
reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

- $TF = \frac{\#\ 'word'\ in\ doc}{tot\ \#\ words\ in\ doc}$
- $IDF = log\left(\frac{\#\ docs\ with\ 'word'\ in\ it}{tot\ docs\ in\ collection}\right)$

The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q27. Why we generally use Soft-max (or sigmoid) non-linearity function as last operation in-network? Why RELU in an inner layer?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).
Then the i-eth component of soft-max(x) is:



It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.
RELU because it avoids the vanishing gradient descent issue.

Data Analysis

Q1. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:
• Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
• R is more suitable for machine learning than just text analysis.
• Python performs faster for all types of text analytics.

Q2. How does data cleaning play a vital role in the analysis?

Data cleaning can help in analysis because:
• Cleaning data from multiple sources helps transform it into a format that data analysts or data scientists can work with.
• Data Cleaning helps increase the accuracy of the model in machine learning.
• It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
• It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

Q3. Differentiate between univariate, bivariate and multivariate analysis.

Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on one variable involved at a given point of time. For example, the pie charts of sales based on territory

involve only one variable and can the analysis can be referred to as univariate analysis.
The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.
Multivariate analysis deals with the study of more than two variables to understand the effect of variables
on the responses.

Q4. Explain Star Schema.

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes
star schemas involve several layers of summarization to recover information faster.

Q5. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample
where each sampling unit is a collection or cluster of elements.
For example, a researcher wants to survey the academic performance of high school students in Japan. He
can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Q6. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the
list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Q7. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the
eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.
Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the
factor by which the compression occurs.

Q8. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are
• False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
• False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.
Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom
they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on
sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not
actually purchased anything but are marked as having made $10,000 worth of purchase.

Q9. Can you cite some examples where a false negative important than a false positive? And vice versa?

Example 1 FN: What if Jury or judge decides to make a criminal go free?
Example 2 FN: Fraud detection.
Example 3 FP: customer voucher use promo evaluation: if many used it and actually if was not true, promo sucks.

Q10. Can you cite some examples where both false positive and false negatives are equally important?

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.
Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

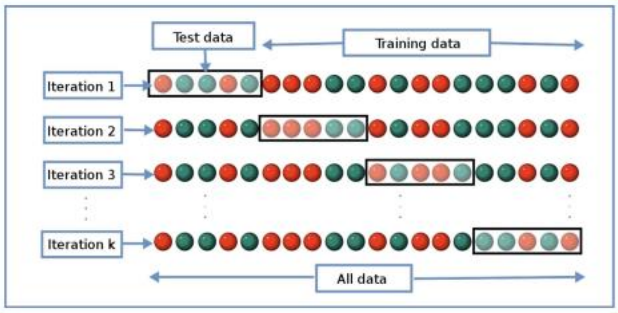
Q11. Can you explain the difference between a Validation Set and a Test Set?

A Training Set:
• to fit the parameters i.e. weights
A Validation set:
• part of the training set
• for parameter selection
• to avoid overfitting
A Test set:
• for testing or evaluating the performance of a trained machine learning model, i.e. evaluating the predictive power and generalization.

Q12. Explain cross-validation.

https://machinelearningmastery.com/k-fold-cross-validation/
Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Mainly used in backgrounds where the objective is forecast, and one wants to estimate how accurately a model will accomplish in practice.
Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to
perform in general when used to make predictions on data not used during the training of the model.
It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.
The general procedure is as follows:
1. Shuffle the dataset randomly.

2. Split the dataset into k groups
3. For each unique group:
a. Take the group as a hold out or test data set
b. Take the remaining groups as a training data set
c. Fit a model on the training set and evaluate it on the test set
d. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores



There is an alternative in Scikit-Learn called Stratified k fold, in which the split is shuffled to make it sure you have a representative sample of each class and a k fold in which you may not have the assurance of it (not good with a very unbalanced dataset).

Machine Learning

Q1. What is Machine Learning?

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. You select a model to train and then manually perform feature extraction. Used to devise complex models and algorithms that lend themselves
to a prediction which in commercial use is known as predictive analytics.

Q2. What is Supervised Learning?

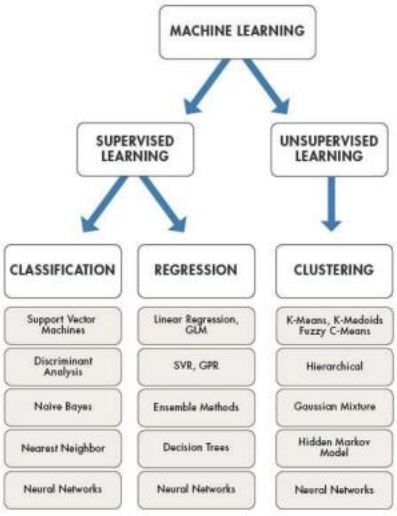
Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.
Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural Networks
E.g. If you built a fruit classifier, the labels will be “this is an orange, this is an apple and this is a banana”, based on showing the classifier examples of apples, oranges and bananas.

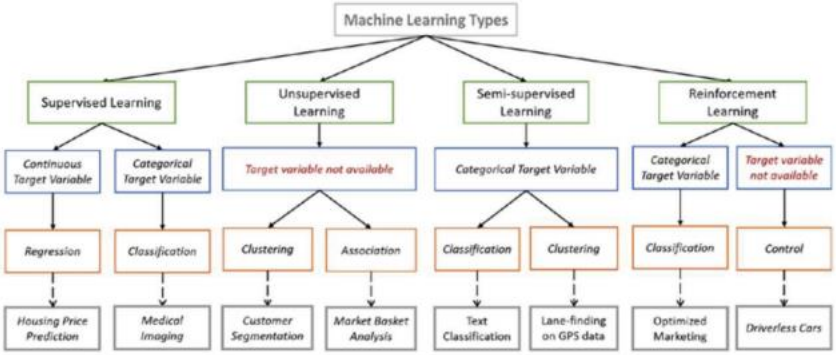
Q3. What is Unsupervised learning?

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.
Algorithms: Clustering, Anomaly Detection, Neural Networks and Latent Variable Models
E.g. In the same example, a fruit clustering will categorize as “fruits with soft skin and lots of dimples”, “fruits with shiny hard skin” and “elongated yellow fruits”.

Q4. What are the various algorithms?

There are various algorithms. Here is a list.





Q5. What is 'Naive' in a Naive Bayes?

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector $x_1$through x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive conditional independence assumption that each x_i is independent: for all i, this relationship is simplified to:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(y|x_i)$; the former is then the relative frequency of class y in the training set.

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

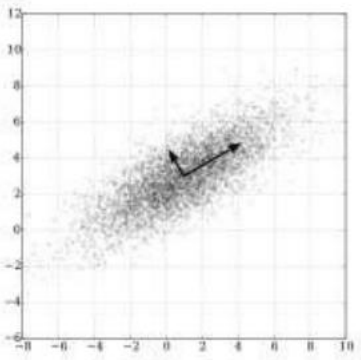
$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(y|x_i)$: can be Bernoulli, Binomial, Gaussian, and so on.

Q6. What is PCA? When do you use it?

Principal component analysis (PCA) is a statistical method used in Machine Learning. It consists in projecting data in a higher dimensional space into a lower dimensional space by maximizing the variance of each dimension.

The process works as following. We define a matrix A with □ rows (the single observations of a dataset – in a tabular format, each single row) and □ columns, our features. For this matrix we construct a variable space with as many dimensions as there are features. Each feature represents one coordinate axis. For each feature, the length has been standardized according to a scaling criterion, normally by scaling to unit variance. It is determinant to scale the features to a common scale, otherwise the features with a greater magnitude will weigh more in determining the principal components. Once plotted all the observations and computed the mean of each variable, that mean will be represented by a point in the center of our plot (the center of gravity). Then, we subtract each observation with the mean, shifting the coordinate system with the center in the origin. The best fitting line resulting is the line that best accounts for the shape of the point swarm. It represents the maximum variance direction in the data. Each observation may be projected onto this line in order to get a coordinate value along the PC-line. This value is known as a score. The next best-fitting line can be similarly chosen from directions perpendicular to the first. Repeating this process yields an orthogonal basis in which different individual dimensions of the data are uncorrelated. These basis vectors are called principal components.



PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used
to visualize genetic distance and relatedness between populations.

Q7. Explain SVM algorithm in detail.

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of supportvector
machines, a data point is viewed as a p-dimensional vector (a list of □ numbers), and we want to know whether we can separate such points with a (Ꝑ − 1)-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So, we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If
such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines
is known as a maximum-margin classifier; or equivalently, the perceptron of optimal stability. The best hyper plane that divides the data is H_3.

We have n data $(x_1, y_1), \dots, (x_n, y_n)$ and p different features $x_i = (x_i^1, \dots, x_i^p)$ and y_i is either 1 or -1.

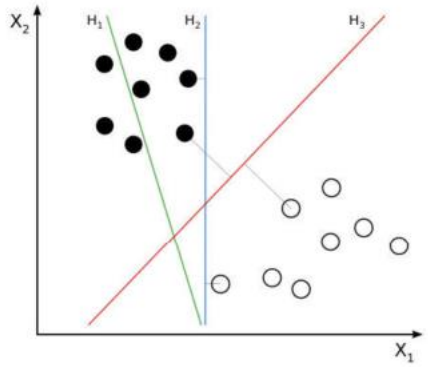
The equation of the hyperplane H_3 is as the set of points x satisfying:

$$w \cdot x - b = 0$$

where w is the (not necessarily normalized) normal vector to the hyperplane. The parameter $\frac{b}{\|w\|}$ determines the offset of the hyperplane from the origin along the normal vector w.

So, for each i, either x_i is in the hyperplane of 1 or -1. Basically, x_i satisfies:

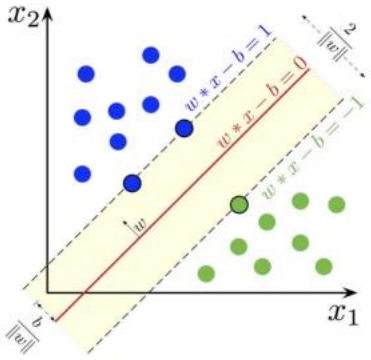
$$w \cdot x_i - b \geq 1 \quad or \quad w \cdot x_i - b \leq -1$$



• SVMs are helpful in text and hypertext categorization, as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings. Some methods for shallow semantic parsing are based on support vector machines.
• Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.
• Classification of satellite data like SAR data using supervised SVM.
• Hand-written characters can be recognized using SVM.

Q8. What are the support vectors in SVM?

In the diagram, we see that the sketched lines mark the distance from the classifier (the hyper plane) to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.



To extend SVM to cases in which the data are not linearly separable, we introduce the hinge loss function,

$$\max(0, 1 - y_i(w \cdot x_i - b))$$

This function is zero if x lies on the correct side of the margin. For data on the wrong side of the margin, the function's value is proportional to the distance from the margin.

Q9. What are the different kernels in SVM?

There are four types of kernels in SVM.

1. LinearKernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

Q10. What are the most known ensemble algorithms?

The most popular trees are: AdaBoost, Random Forest, and eXtreme Gradient Boosting (XGBoost).
AdaBoost is best used in a dataset with low noise, when computational complexity or timeliness of results
is not a main concern and when there are not enough resources for broader hyperparameter tuning due to lack of time and knowledge of the user.
Random forests should not be used when dealing with time series data or any other data where lookahead
bias should be avoided, and the order and continuity of the samples need to be ensured. This algorithm can handle noise relatively well, but more knowledge from the user is required to adequately tune the algorithm compared to AdaBoost.
The main advantages of XGBoost is its lightning speed compared to other algorithms, such as AdaBoost, and its regularization parameter that successfully reduces variance. But even aside from the regularization
parameter, this algorithm leverages a learning rate (shrinkage) and subsamples from the features like random forests, which increases its ability to generalize even further. However, XGBoost is more difficult to understand, visualize and to tune compared to AdaBoost and random forests. There is a multitude of hyperparameters that can be tuned to increase performance.

Q11. Explain Decision Tree algorithm in detail.

A decision tree is a supervised machine learning algorithm mainly used for Regression and Classification. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree can handle both categorical and numerical data. The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures.
Some techniques, often called ensemble methods, construct more than one decision tree:

- Boosted trees Incrementally building an ensemble by training each new instance to emphasize
the training instances previously mis-modeled. A typical example is AdaBoost. These can be used
for regression-type and classification-type problems.
- Bootstrap aggregated (or bagged) decision trees, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for
a consensus prediction.
    - A random forest classifier is a specific type of bootstrap aggregating.

• Rotation forest – in which every decision tree is trained by first applying principal component
analysis (PCA) on a random subset of the input features.
A special case of a decision tree is a decision list, which is a one-sided decision tree, so that every internal
node has exactly 1 leaf node and exactly 1 internal node as a child (except for the bottommost node,
whose only child is a single leaf node). While less expressive, decision lists are arguably easier to
understand than general decision trees due to their added sparsity, permit non-greedy learning methods
and monotonic constraints to be imposed.
Notable decision tree algorithms include:
• ID3 (Iterative Dichotomiser 3)
• C4.5 (successor of ID3)
• CART (Classification and Regression Tree)
• Chi-square automatic interaction detection (CHAID). Performs multi-level splits when computing
classification trees.
• MARS: extends decision trees to handle numerical data better.
• Conditional Inference Trees. Statistics-based approach that uses non-parametric tests as splitting
criteria, corrected for multiple testing to avoid overfitting. This approach results in unbiased
predictor selection and does not require pruning.

Q12. What are Entropy and Information gain in Decision tree algorithm?

There are a lot of algorithms which are employed to build a decision tree, ID3 (Iterative Dichotomiser 3),
C4.5, C5.0, CART (Classification and Regression Trees) to name a few but at their core all of them tell us
what questions to ask and when.
The below table has color and diameter of a fruit and the label tells the name of the fruit. How do we build
a decision tree to classify the fruits?

Color	Diameter	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Here is how we will build the tree. We will start with a node which will ask a true or false question to split
the data into two. The two resulting nodes will each ask a true or false question again to split the data
further and so on.
There are 2 main things to consider with the above approach:
• Which is the best question to ask at each node
• When do we stop splitting the data further?
Let's start building the tree with the first or the topmost node. There is a list of possible questions which
can be asked. The first node can ask the following questions:
• Is the color green?
• Is the color yellow?
• Is the color red?
• Is the diameter ≥ 3?
• Is the diameter ≥ 1?
Of these possible set of questions, which one is the best to ask so that our data is split into two sets after
the first node? Remember we are trying to split or classify our data into separate classes. Our question
should be such that our data is partitioned into as unmixed or pure classes as possible. An impure set or
class here refers to one which has many different types of objects for example if we ask the question for
the above data, "Is the color green?" our data will be split into two sets one of which will be pure the other
will have a mixed set of labels. If we assign a label to a mixed set, we have higher chances of being
incorrect. But how do we measure this impurity?

Color	Diameter	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Gini Impurity and Information Gain – CART

CART (Classification and Regression Trees) → uses **Gini Index (Classification)** as metric.

The Gini Impurity (GI) metric measures the homogeneity of a set of items. The lowest possible value of GI
is 0.0. The maximum value of GI depends on the particular problem being investigated but gets close to
1.0.
Suppose for example you have 12 items — apples, grapes, lemons. If there are 0 apples, 0 grapes, 12 lemons,
then you have minimal impurity (this is good for decision trees) and GI = 0.0. But if you have 4 apples, 4
grapes, 4 lemons, you have maximum impurity and it turns out that GI = 0.667.
I'll show example calculations.
Maximum GI: Apples, Grapes, Lemons

```
count = 4 4 4
p = 4/12 4/12 4/12
= 1/3 1/3 1/3

GI = 1 - [ (1/3)^2 + (1/3)^2 + (1/3)^2 ]
= 1 - [ 1/9 + 1/9 + 1/9 ]
= 1 - 1/3
= 2/3
= 0.667
```

When the number of items is evenly distributed, as in the example above, you have maximum GI but the
exact value depends on how many items there are. A bit less than maximum GI:

```
count = 3 3 6
p = 3/12 3/12 6/12
= 1/4 1/4 1/2

GI = 1 - [ (1/4)^2 + (1/4)^2 + (1/2)^2 ]
= 1 - [ 1/16 + 1/16 + 1/4 ]
= 1 - 6/16
= 10/16
= 0.625
```

In the example above, the items are not quite evenly distributed, and the GI is slightly less (which is better
when used for decision trees). Minimum GI:

```
count = 0 12 0
p = 0/12 12/12 0/12
= 0 1 0

GI = 1 - [ 0^2 + 1^2 + 0^2 ]
= 1 - [ 0 + 1 + 0 ]
= 1 - 1
= 0.00
```

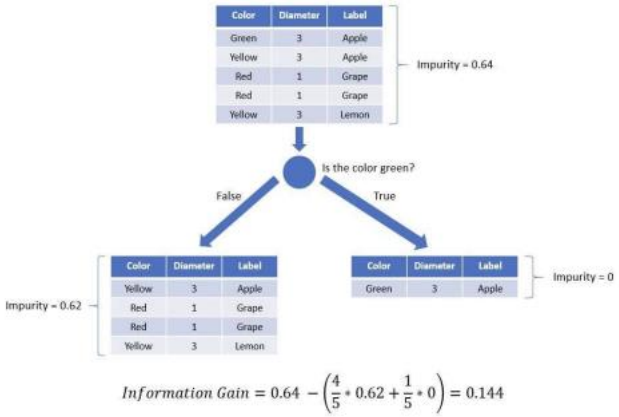
J classes $\{1, 2, \ldots J\}$, p_i: fraction of elements of class i:

Gini impurity: $I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$

The **Gini index** is not at all the same as a different metric called the **Gini coefficient.** The Gini impurity
metric can be used when creating a decision tree but there are alternatives, including
Entropy
Information gain. The advantage of GI is its simplicity.

Information Gain

Information gain is another metric which tells us how much a question unmixes the labels at a
node. **"Mathematically it is just a difference between impurity values before splitting the data at a node and the weighted average of the impurity after the split"**. For instance, if we go back to our data of apples, lemons and grapes and ask the question "Is the color Green?"



$$Information\ Gain = 0.64 - \left(\frac{4}{5} * 0.62 + \frac{1}{5} * 0\right) = 0.144$$

The information gain by asking this question is 0.144. Similarly, we can ask another question from the set

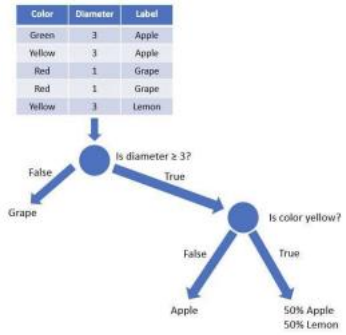
of possible questions split the data and compute information gain. This is also called (Recursive Binary Splitting).

Question	Information Gain
Is the color green?	0.14
Is diameter ≥ 3?	0.37
Is the color yellow?	0.17
Is the color red?	0.37
Is diameter ≥ 1?	0

The question where we have the highest information gain “Is diameter ≥ 3?” is the best question to ask.

Note that the information gain is same for the question “Is the color red?” we just picked the first one at random.

Repeating the same method at the child node we can complete the tree. Note that no further questions can be asked which would increase the information gain.

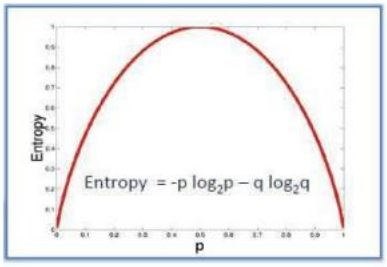


Also note that the rightmost leaf which says 50% Apple & 50% lemon means that this class cannot be divided further, and this branch can tell an apple or a lemon with 50% probability. For the grape and apple branches we stop asking further questions since the Gini Impurity is 0 for those.

Entropy and Information Gain – ID3

ID3 (Iterative Dichotomiser 3) → uses **Entropy function** and Information gain as metrics.

If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has
entropy of one.



Entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

b) Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = P(Sunny)*E(3,2) + P(Overcast)*E(4,0) + P(Rainy)*E(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute.

Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e.,

the most homogeneous branches).

Step 1: Calculate entropy of the target.

Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated.

Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from

the entropy before the split. The result is the Information Gain or decrease in entropy.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

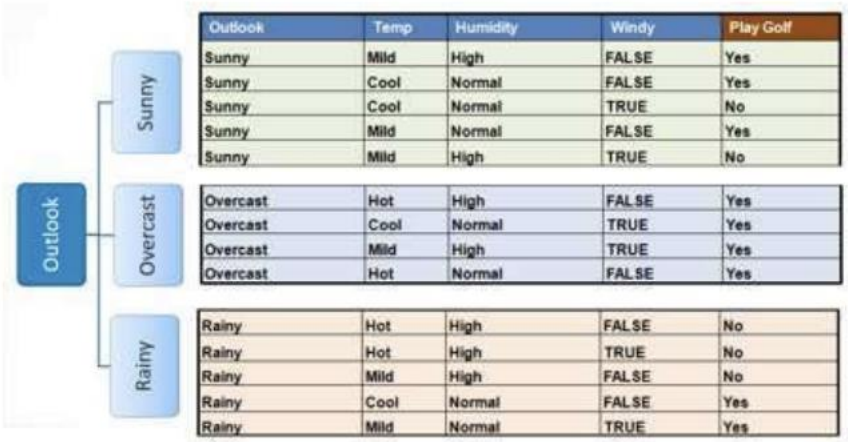
$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its
branches and repeat the same process on every branch.

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			



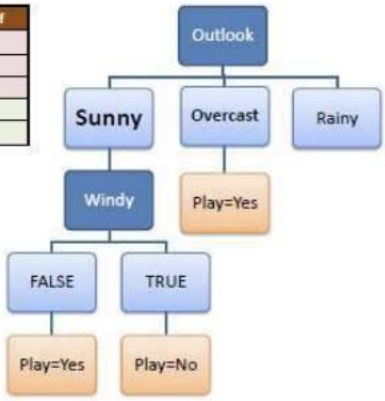
Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Step 4a: A branch with entropy of 0 is a leaf node.

Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes

Step 4b: A branch with entropy more than 0 needs further splitting.

Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

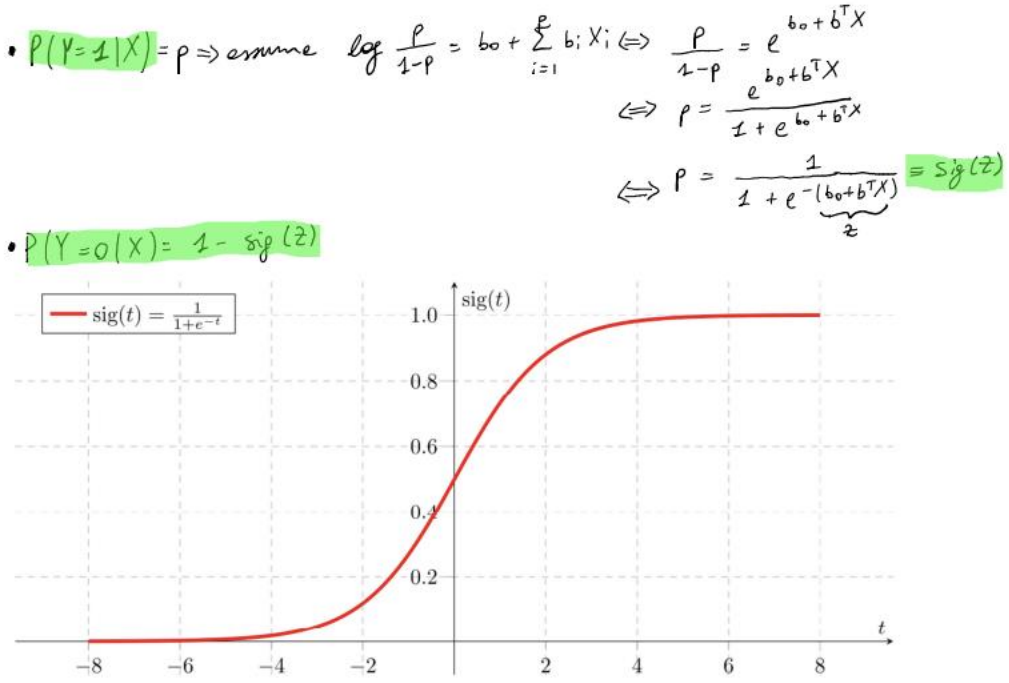
Q13. What is pruning in Decision Tree?

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees
by removing sections of the tree that provide little power to classify instances. So, when we remove subnodes
of a decision node, this process is called pruning or opposite process of splitting.

Q14. What is logistic regression? State an example when you have used logistic regression recently.

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from
a linear combination of predictor variables. Since we are interested in a probability outcome, a line does
not fit the model. Logistic Regression is a classification algorithm that works by trying to learn a function

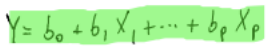
that approximates P(Y|X) . It makes the central assumption that P(X|Y) can be approximated as a
sigmoid function applied to a linear combination of input features.

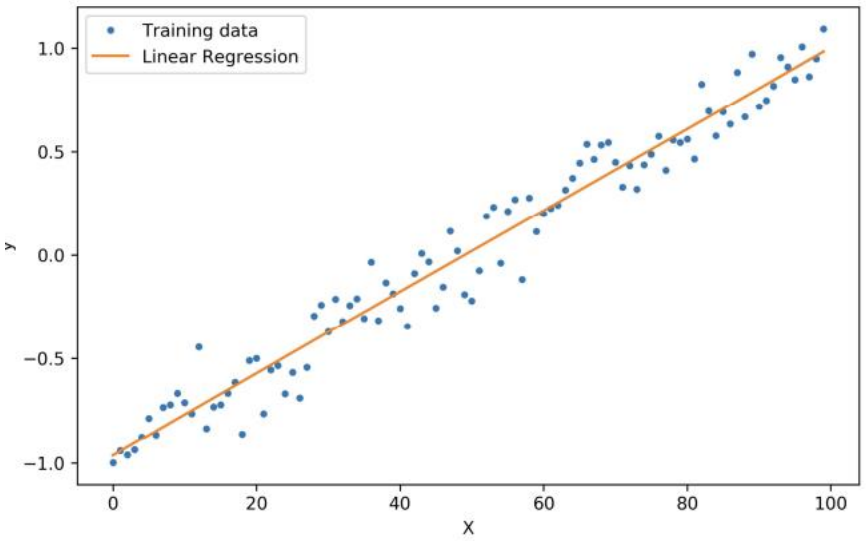


For example, if you want to predict whether a particular political leader will win the election or not. In this
case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the
amount of money spent for election campaigning of a particular candidate, the amount of time spent in
campaigning, etc.

Q15. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a
second variable X. X is referred to as the predictor variable and Y as the criterion variable.





Q16. What Are the Drawbacks of the Linear Model?

Some drawbacks of the linear model are:

- The assumption of linearity of the model
- It can't be used for count outcomes or binary outcomes.
- There are overfitting or underfitting problems that it can't solve.

Q17. What is the difference between Regression and classification ML techniques?

Both Regression and classification machine learning techniques come under Supervised machine learning
algorithms. In Supervised machine learning algorithm, we have to train the model using labelled data set,
while training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from
input to output. If our labels are discrete values then it will a classification problem, but if our labels are continuous values then it will be a regression problem.

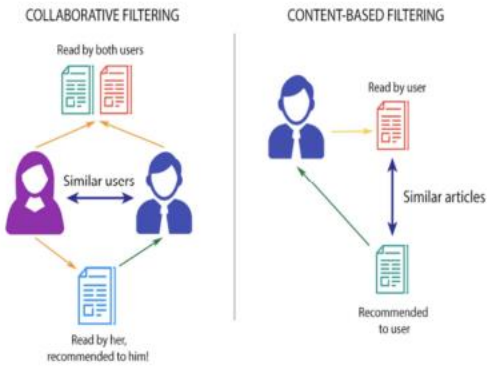
Q18. What are Recommender Systems?

Recommender Systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.
Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in ecommerce
sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q19. What is Collaborative filtering? And a content based?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents. Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user.

It looks at the items they like (usually based on rating) and combines them to create a ranked list of suggestions. Similar users are those with similar rating and on the based on that they get recommendations. In content based, we look only at the item level, recommending on similar items sold.



An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others' ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay &
Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q20. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number
of outlier values is few then they can be assessed individually but for a large number of outliers, the values
can be substituted with either the 99th or the 1st percentile values.
All extreme values are not outlier values. The most common ways to treat outlier values:
1. Change it with a mean or median
2. Standardize the feature, changing the distribution but smoothing the outliers
3. Log transform the feature (with many outliers)
4. Drop the value
5. First/third quartile value if more than 2

Q21. What are the various steps involved in an analytics project?

The following are the various steps involved in an analytics project:
1. Understand the Business problem
2. Explore the data and become familiar with it
3. Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

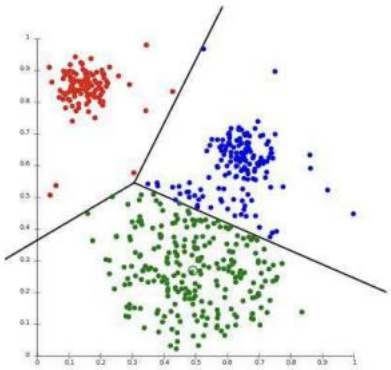
Q22. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values
(imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.
If 80% of the values for a variable are missing, then you can answer that you would be dropping the variable instead of treating the missing values.

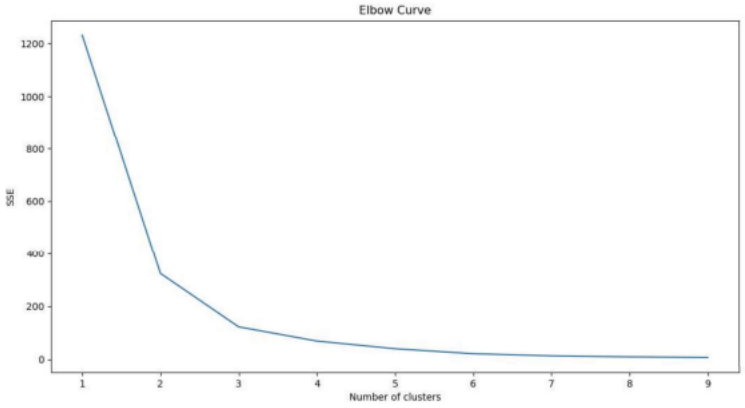
Q23. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to K-Means clustering
where "K" defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other, but the groups are different from each other. For example, the following image shows three different groups.

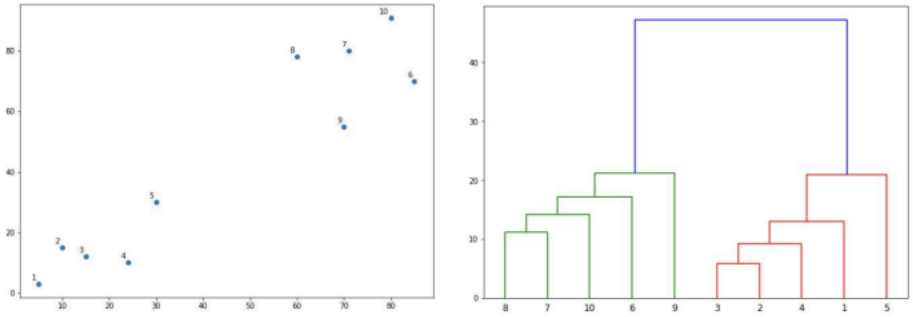


Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS (as the sum of the squared distance between each member of the cluster and its centroid) for a range of number of clusters, you will get the plot shown below.

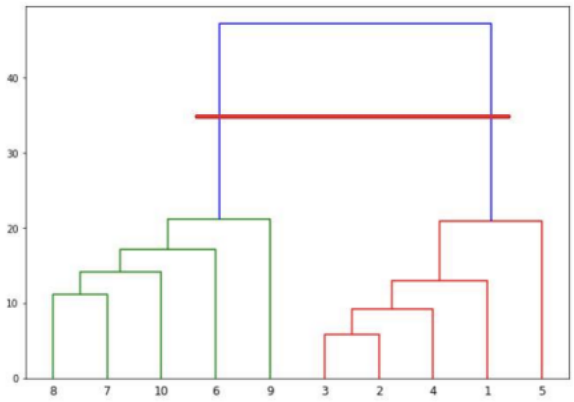
- The Graph is generally known as Elbow Curve.
- Red circled a point in above graph i.e. Number of Cluster = 3 is the point after which you don't see any decrement in WSS.
- This point is known as the bending point and taken as K in K – Means.



This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

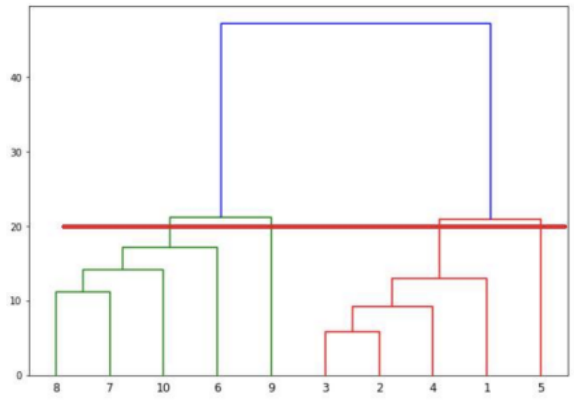


The algorithm starts by finding the two points that are closest to each other on the basis of Euclidean distance. If we look back at Graph1, we can see that points 2 and 3 are closest to each other while points 7 and 8 are closes to each other. Therefore a cluster will be formed between these two points first. In Graph2, you can see that the dendograms have been created joining points 2 with 3, and 8 with 7. The vertical height of the dendogram shows the Euclidean distances between points. From Graph2, it can be seen that Euclidean distance between points 8 and 7 is greater than the distance between point 2 and 3. The next step is to join the cluster formed by joining two points to the next nearest cluster or point which
in turn results in another cluster. If you look at Graph1, point 4 is closest to cluster of point 2 and 3, therefore in Graph2 dendrogram is generated by joining point 4 with dendrogram of point 2 and 3. This process continues until all the points are joined together to form one big cluster.
Once one big cluster is formed, the longest vertical distance without any horizontal line passing through it is selected and a horizontal line is drawn through it. The number of vertical lines this newly created horizontal line passes is equal to number of clusters. Take a look at the following plot:



We can see that the largest vertical distance without any horizontal line passing through it is represented
by blue line. So we draw a new horizontal red line that passes through the blue line. Since it crosses the blue line at two points, therefore the number of clusters will be 2. Basically the horizontal line is a threshold, which defines the minimum distance required to be a separate cluster. If we draw a line further
down, the threshold required to be a new cluster will be decreased and more clusters will be formed as

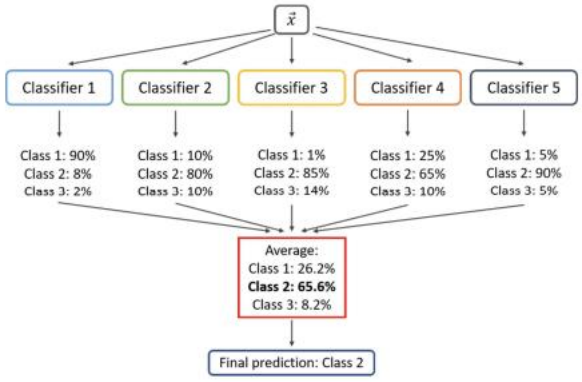
see in the image below:



In the above plot, the horizontal line passes through four vertical lines resulting in four clusters: cluster of points 6,7,8 and 10, cluster of points 3,2,4 and points 9 and 5 will be treated as single point clusters.

Q24. What is Ensemble Learning?

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Ensembles are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. Each classifier, individually, is a “weak learner,” while all the classifiers taken together are a “strong learner”.



Q25. Describe in brief any type of Ensemble Learning.

Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

Bagging

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the
predictions. In generalized bagging, you can use different learners on different population. As you expect
this helps us to reduce the variance error.

Pros

Ø Bagging method helps when we face variance or overfitting in the model. It provides an environment to deal with variance by using N learners of same size on same algorithm.
Ø During the sampling of train data, there are many observations which overlaps. So, the combination of these learners helps in overcoming the high variance.
Ø Bagging uses Bootstrap sampling method (Bootstrapping is any test or metric that uses random sampling with replacement and falls under the broader class of resampling methods.)

Cons

Ø Bagging is not helpful in case of bias or underfitting in the data.
Ø Bagging ignores the value with the highest and the lowest result which may have a wide difference and provides an average result.

Boosting

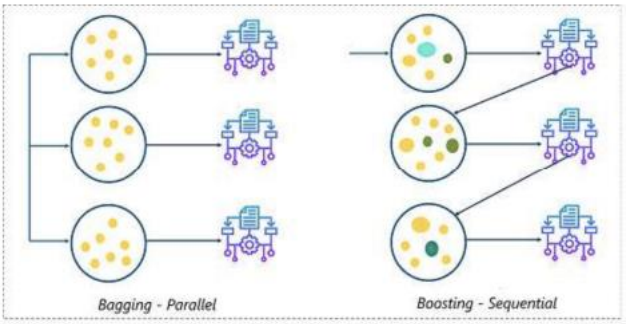
Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation
and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However,
they may over fit on the training data.

Pros

Ø Boosting technique takes care of the weightage of the higher accuracy sample and lower accuracy sample and then gives the combined results.
Ø Net error is evaluated in each learning steps. It works good with interactions.
Ø Boosting technique helps when we are dealing with bias or underfitting in the data set.
Ø Multiple boosting techniques are available. For example: AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost

Cons

Ø Boosting technique often ignores overfitting or variance issues in the data set.
Ø It increases the complexity of the classification.
Ø Time and computation can be a bit expensive.



There are multiple areas where Bagging and Boosting technique is used to boost the accuracy.

- Banking: Loan defaulter prediction, fraud transaction
- Credit risks
- Kaggle competitions
- Fraud detection
- Recommender system for Netflix
- Malware

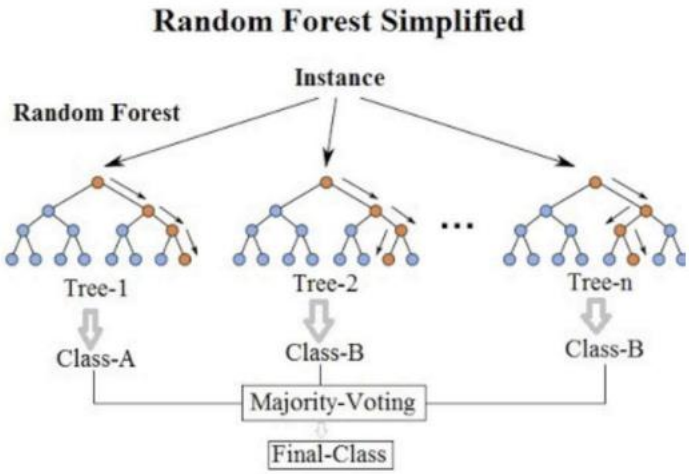
• Wildlife conservations and so on.

Q26. What is a Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing:

• regression
• classification
• dimensionality reduction
• treat missing values
• outlier values

It is a type of ensemble learning method, where a group of weak models combine to form a powerful model. The random forest starts with a standard machine learning technique called a “decision tree” which, in ensemble terms, corresponds to our weak learner. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets.



In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes (Over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

Q27. How Do You Work Towards a Random Forest?

The underlying principle of this technique is that several weak learners combined to provide a keen learner. Here is how such a system is trained for some number of trees T:

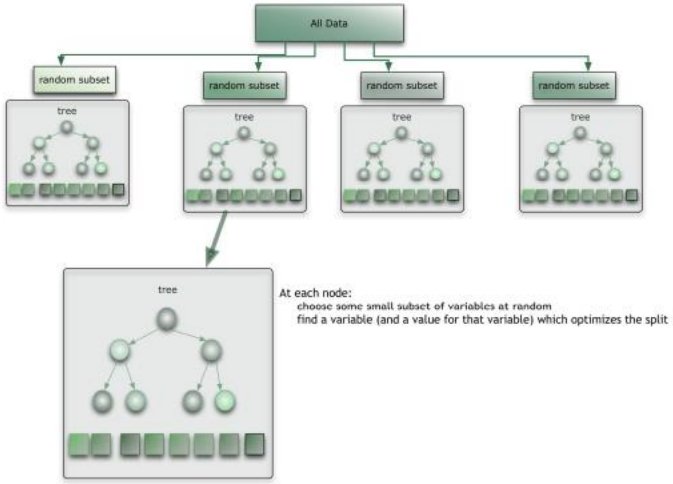
1. Sample N cases at random with replacement to create a subset of the data. The subset should be about 66% of the total set.
2. At each node:

a. For some number m (see below), m predictor variables are selected at random from all the predictor variables.
b. The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
c. At the next node, choose another m variables at random from all predictor variables and do the same.

Depending upon the value of m, there are three slightly different systems:

• Random splitter selection: $m = 1$

- Breiman's bagger: m = total number of predictor variables (p)
- Random forest: m ≪ number of predictor variables.
    - Brieman suggests three possible values for m: $\frac{1}{2}\sqrt{p}$, $\sqrt{p}$, $2\sqrt{p}$



When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.
Note that:

Ø With a large number of predictors (p ≫ 0), the eligible predictor set (m) will be quite different from node to node.

Ø The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.

Ø As m goes down, both inter-tree correlation and the strength of individual trees go down. So some optimal value of m must be discovered.

Ø Strengths: Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data.

Ø Weaknesses: Random Forest used for regression cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon your own data set.

Q28. What cross-validation technique would you use on a time series data set?

Instead of using k-fold cross-validation, you should be aware of the fact that a time series is not randomly
distributed data — It is inherently ordered by chronological order.
In case of time series data, you should use techniques like forward=chaining — Where you will be model on past data then look at forward-facing data.
fold 1: training[1], test[2]
fold 2: training[1 2], test[3]
fold 3: training[1 2 3], test[4]
fold 4: training[1 2 3 4], test[5]

Q29. What is a Box-Cox Transformation?

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary

least squares regression. The residuals could either curve as the prediction increases or follow the skewed
distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box-Cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a Box-Cox transformation means that you can run a broader number of tests.
A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named
after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed
the technique.

Q30. How Regularly Must an Algorithm be Updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
- There is a case of non-stationarity (mean, variance change over the time)
- The algorithm underperforms/results lack accuracy

Q31. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all, you have to ask which ML model you want to train.
For Neural networks: Batch size with Numpy array will work. Steps:
1. Load the whole data in the Numpy array. Numpy array has a property to create a mapping of the complete data set, it doesn't load complete data set in memory.
2. You can pass an index to Numpy array to get required data.
3. Use this data to pass to the Neural network.
4. Have a small batch size.

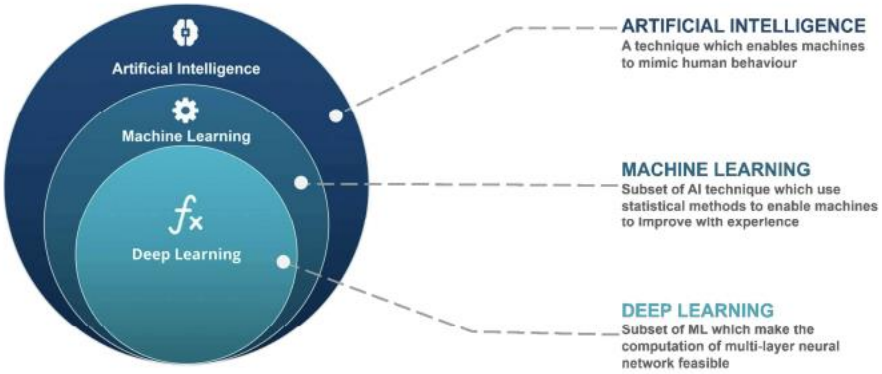
For SVM: Partial fit will work. Steps:
1. Divide one big data set in small size data sets.
2. Use a partial fit method of SVM, it requires a subset of the complete data set.
3. Repeat step 2 for other subsets.

However, you could actually face such an issue in reality. So, you could check out the best laptop for Machine Learning to prevent that. Having said that, let's move on to some questions on deep learning.

Deep Learning

Q1. What do you mean by Deep Learning?

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of
the neurons in the human brain.

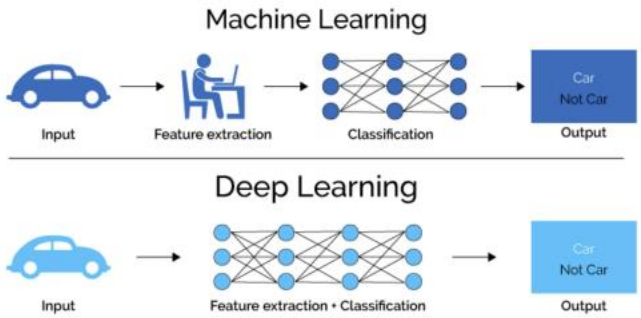


Q2. What is the difference between machine learning and deep learning?

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorized in the following four categories.

1. Supervised machine learning,
2. Semi-supervised machine learning,
3. Unsupervised machine learning,
4. Reinforcement learning.

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.



• The main difference between deep learning and machine learning is due to the way data is presented in the system. Machine learning algorithms almost always require structured data, while deep learning networks rely on layers of ANN (artificial neural networks).

• Machine learning algorithms are designed to “learn” to act by understanding labeled data and then use it to produce new results with more datasets. However, when the result is incorrect, there is a need to “teach them”. Because machine learning algorithms require bulleted data, they are not suitable for solving complex queries that involve a huge amount of data.

• Deep learning networks do not require human intervention, as multilevel layers in neural networks place data in a hierarchy of different concepts, which ultimately learn from their own mistakes. However, even they can be wrong if the data quality is not good enough.

• Data decides everything. It is the quality of the data that ultimately determines the quality of the result.

• Both of these subsets of AI are somehow connected to data, which makes it possible to represent a certain form of “intelligence.” However, you should be aware that deep learning requires much more data than a traditional machine learning algorithm. The reason for this is that deep learning networks can identify different elements in neural network layers only when more than a million data points interact. Machine learning algorithms, on the other hand, are capable of learning by

pre-programmed criteria.

Q3. What, in your opinion, is the reason for the popularity of Deep Learning in recent times?

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

• The increase in the amount of data generated through various sources

• The growth in hardware resources required to run these models

GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously.

Q4. What is reinforcement learning?

Reinforcement Learning allows to take actions to max cumulative reward. It learns by trial and error through reward/penalty system. Environment rewards agent so by time agent makes better decisions. Ex: robot=agent, maze=environment. Used for complex tasks (self-driving cars, game AI).

RL is a series of time steps in a Markov Decision Process:

1. Environment: space in which RL operates
2. State: data related to past action RL took
3. Action: action taken
4. Reward: number taken by agent after last action
5. Observation: data related to environment: can be visible or partially shadowed

Q5. What are Artificial Neural Networks?

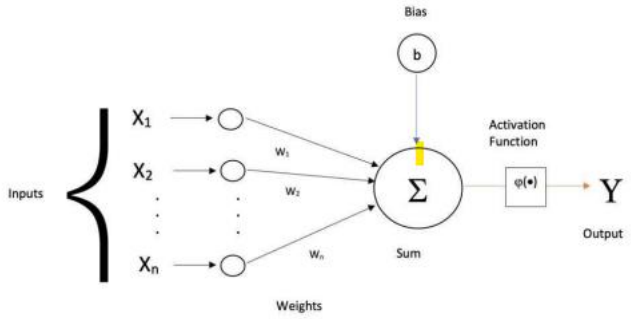
Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They

are inspired by biological neural networks. Neural Networks can adapt to changing the input, so the network generates the best possible result without needing to redesign the output criteria.

Q6. Describe the structure of Artificial Neural Networks?

Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs

which get processed with weighted sums and Bias, with the help of Activation Functions.



Q7. How Are Weights Initialized in a Network?

There are two methods here: we can either initialize the weights to zero or assign them randomly. Initializing all weights to 0: This makes your model similar to a linear model. All the neurons and every layer perform the same operation, giving the same output and making the deep net useless.

Initializing all weights randomly: Here, the weights are assigned randomly by initializing them very close to 0. It gives better accuracy to the model since every neuron performs different computations. This is the

most commonly used method.

Q8. What Is the Cost Function?

Also referred to as “loss” or “error,” cost function is a measure to evaluate how good your model’s performance is. It’s used to compute the error of the output layer during backpropagation. We push that
error backwards through the neural network and use that during the different training functions.
The most known one is the mean sum of squared errors.

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} \underbrace{(y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i)^2}_{\text{actual value}}$$

$$\hat{y}_i = \phi(\sum(w_i x_i) + b)$$

Q9. What Are Hyperparameters?

With neural networks, you’re usually working with hyperparameters once the data is formatted correctly.
A hyperparameter is a parameter whose value is set before the learning process begins. It determines how a network is trained and the structure of the network (such as the number of hidden units, the learning rate, epochs, batches, etc.).

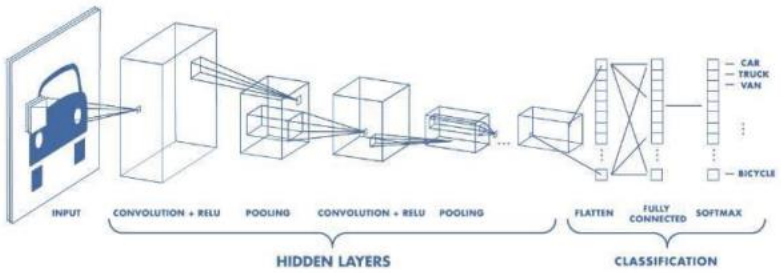
Q10. What Will Happen If the Learning Rate Is Set inaccurately (Too Low or Too High)?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point.
If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is
too chaotic for the network to train).

Q11. What Is The Difference Between Epoch, Batch, and Iteration in Deep Learning?

- Epoch – Represents one iteration over the entire dataset (everything put into the training model).
- Batch – Refers to when we cannot pass the entire dataset into the neural network at once, so we divide the dataset into several batches.
- Iteration – if we have 10,000 images as data and a batch size of 200. then an epoch should run 50 iterations (10,000 divided by 50).

Q12. What Are the Different Layers on CNN?



The Convolutional neural networks are regularized versions of multilayer perceptron (MLP). They were developed based on the working of the neurons of the animal visual cortex.

Let's say we have a color image in JPG form and its size is 480 x 480. The representative array will be 480 x 480 x 3. Each of these numbers is given a value from 0 to 255 which describes the pixel intensity at that
point. RGB intensity values of the image are visualized by the computer for processing.

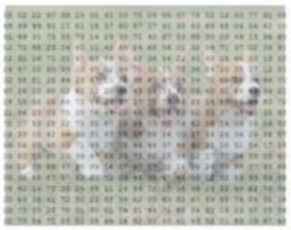
The objective of using the CNN:

The idea is that you give the computer this array of numbers and it will output numbers that describe the
probability of the image being a certain class (.80 for a cat, .15 for a dog, .05 for a bird, etc.). It works similar to how our brain works. When we look at a picture of a dog, we can classify it as such if the picture
has identifiable features such as paws or 4 legs. In a similar way, the computer is able to perform image classification by looking for low-level features such as edges and curves and then building up to more abstract concepts through a series of convolutional layers. The computer uses low-level features obtained
at the initial levels to generate high-level features such as paws or eyes to identify the object.
There are four layers in CNN:
1. Convolutional Layer – the layer that performs a convolutional operation, creating several smaller picture windows to go over the data.
2. Activation Layer (ReLU Layer) – it brings non-linearity to the network and converts all the negative pixels to zero. The output is a rectified feature map. It follows each convolutional layer.
3. Pooling Layer – pooling is a down-sampling operation that reduces the dimensionality of the feature map. Stride = how much you slide, and you get the max of the □ □ □ matrix
4. Fully Connected Layer – this layer recognizes and classifies the objects in the image.

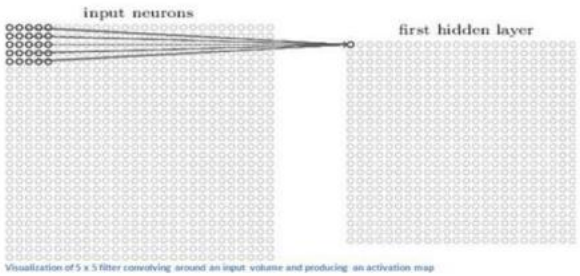
Convolution Operation

First Layer:

1. Input to a convolutional layer
The image is resized to an optimal size and is fed as input to the convolutional layer.
Let us consider the input as 32x32x3 array of pixel values.



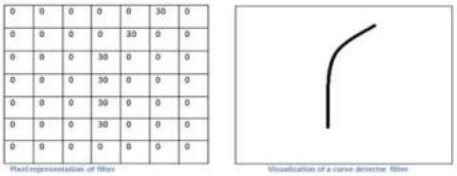
2. There exists a filter or neuron or kernel which lays over some of the pixels of the input image depending on the dimensions of the Kernel size.
Let the dimensions of the kernel of the filter be 5x5x3.



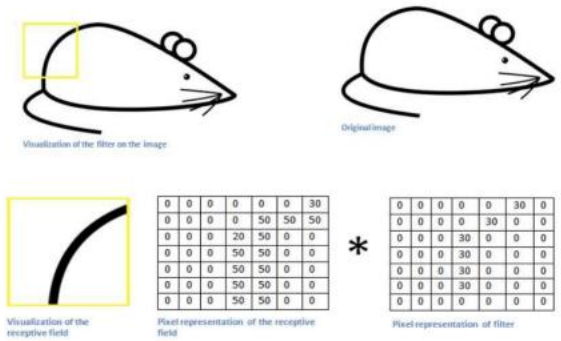
3. The Kernel actually slides over the input image; thus, it is multiplying the values in the filter with the original pixel values of the image (aka computing element-wise multiplications).
The multiplications are summed up generating a single number for that particular receptive field and hence for sliding the kernel a total of 784 numbers are mapped to 28x28 array known as the feature map.
**Now if we consider two kernels of the same dimension then the obtained first layer feature map will be (28x28x2).

High-level Perspective

• Let us take a kernel of size (7x7x3) for understanding. Each of the kernels is considered to be a feature identifier, hence say that our filter will be a curve detector.

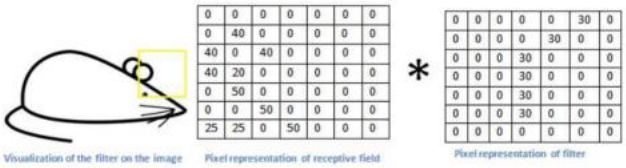


- The original image and the visualization of the kernel on the image.



The sum of the multiplication value that is generated is $= 4 * (50 * 30) + (20 * 30) = 6600$ *(large number).*

- Now when the kernel moves to the other part of the image.

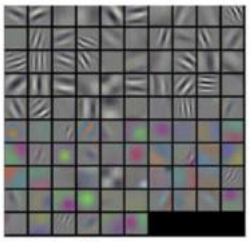


The sum of the multiplication value that is generated is = 0 (small number).

The use of the small and the large value

1. The value is much lower! This is because there wasn't anything in the image section that responded to the curve detector filter. Remember, the output of this convolution layer is an activation map. So, in the simple case of a one filter convolution (and if that filter is a curve detector), the activation map will show the areas in which there at most likely to be curved in the picture.
2. In the previous example, the top-left value of our 26 x 26 x 1 activation map (26 because of the 7x7 filter instead of 5x5) will be 6600. This high value means that it is likely that there is some sort of curve in the input volume that caused the filter to activate. The top right value in our activation map will be 0 because there wasn't anything in the input volume that caused the filter to activate. This is just for one filter.
3. This is just a filter that is going to detect lines that curve outward and to the right. We can have other filters for lines that curve to the left or for straight edges. The more filters, the greater the depth of the activation map, and the more information we have about the input volume.

In the picture, we can see some examples of actual visualizations of the filters of the first conv. layer of a trained network. Nonetheless, the main argument remains the same. The filters on the first layer convolve around the input image and "activate" (or compute high values) when the specific feature it is looking for is in the input volume.



Sequential convolutional layers after the first one

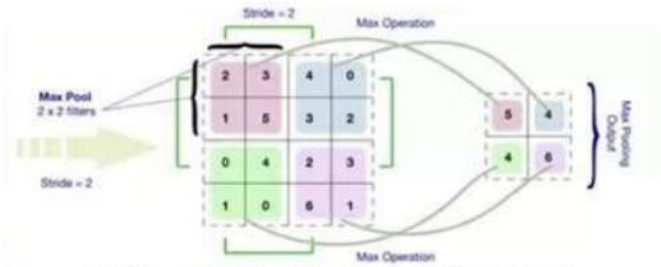
Sequential convolutional layers after the first one

1. When we go through another conv. layer, the output of the first conv. layer becomes the input of the 2nd conv. layer.
2. However, when we're talking about the 2nd conv. layer, the input is the activation map(s) that result from the first layer. So, each layer of the input is basically describing the locations in the original image for where certain low-level features appear.
3. Now when you apply a set of filters on top of that (pass it through the 2nd conv. layer), the output will be activations that represent higher-level features. Types of these features could be semicircles (a combination of a curve and straight edge) or squares (a combination of several straight edges). As you go through the network and go through more convolutional layers, you get activation maps that represent more and more complex features.
4. By the end of the network, you may have some filters that activate when there is handwriting in the image, filters that activate when they see pink objects, etc.

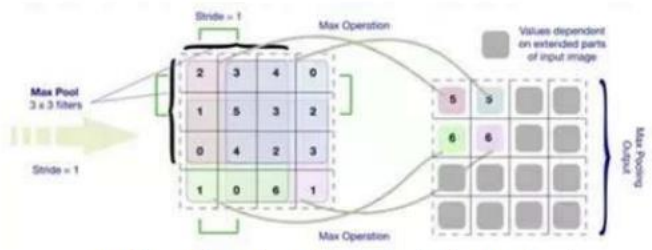
Pooling Operation

It consists in getting the largest number out of a matrix to get the most important number and reduce the dimention.

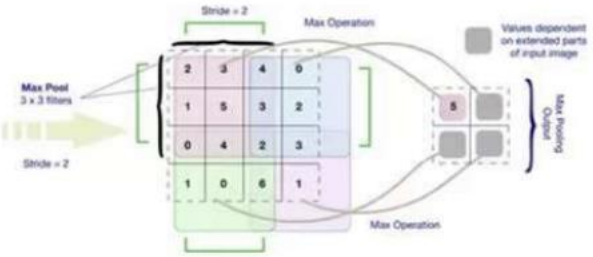
Max Pooling example



2x2 filters with stride = 2 (maximum value) is considered



3x3 filters with stride = 1 (maximum value) is considered



3x3 filters with stride = 2 (maximum value) is considered

Classification

1. Flatten: The pooled matrix is converted to a vector.
2. Fully Connected layer: The way this fully connected layer works is that it looks at the output of the previous layer (which as we remember should represent the activation maps of high-level features) and the number of classes p (10 for digit classification). For example, if the program is predicting that some image is a dog, it will have high values in the activation maps that represent high-level features like a paw or 4 legs, etc. Basically, an FC layer looks at what high level features most strongly correlate to a particular class and has particular weights so that when you compute the products between the weights and the previous layer, you get the correct probabilities for the different classes.
3. Soft-max approach: The output of a fully connected layer is as follows [0 .1 .1 .75 0 0 0 0 0 .05], then this represents a 10% probability that the image is a 1, a 10% probability that the image is a 2, a 75% probability that the image is a 3, and a 5% probability that the image is a 9 (SoftMax approach) for digit classification.

Training

§We know kernels also known as feature identifiers, used for identification of specific features. But how the kernels are initialized with the specific weights or how do the filters know what values to have.

Hence comes the important step of training. The training process is also known as backpropagation, which
is further separated into 4 distinct sections or processes.

- Forward Pass
- Loss Function
- Backward Pass
- Weight Update

The Forward Pass

For the first epoch or iteration of the training the initial kernels of the first convolutional layer are initialized with random values. Thus, after the first iteration output will be something like [.1.1.1.1.1.1.1.1.1.1], which does not give preference to any class as the kernels don't have specific weights.

The Loss Function

The training involves images along with labels, hence the label for the digit 3 will be [0 0 0 1 0 0 0 0 0 0], whereas the output after a first epoch is very different, hence we will calculate loss (MSE — Mean Squared
Error)

$$E_{total} = \sum \frac{1}{2}(target - output)^2$$

The objective is to minimize the loss, which is an optimization problem in calculus. It involves trying to adjust the weights to reduce the loss.

The Backward Pass

It involves determining which weights contributed most to the loss and finding ways to adjust them so that the loss decreases. It is computed using $\frac{\partial L}{\partial W}$ (or $\nabla_w L$), where L is the loss and the W is the weights of the corresponding kernel.

The weights update

This is where the weights of the kernel are updated using the following equation.

$$w = w_i - \eta \frac{dL}{dW}$$

w = Weight
w_i = Initial Weight
η = Learning Rate

Here the Learning Rate is chosen by the programmer. Larger value of the learning rate indicates much larger steps towards optimization of steps and larger time to convolve to an optimized weight.

Testing

Finally, to see whether or not our CNN works, we have a different set of images and labels (can't double dip between training and test!) and pass the images through the CNN. We compare the outputs to the ground truth and see if our network works!

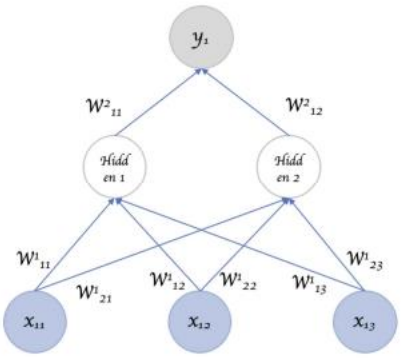
Q13. What Is Pooling on CNN, and How Does It Work?

Pooling is used to reduce the spatial dimensions of a CNN. It performs down-sampling operations to reduce the dimensionality and creates a pooled feature map by sliding a filter matrix over the input matrix.

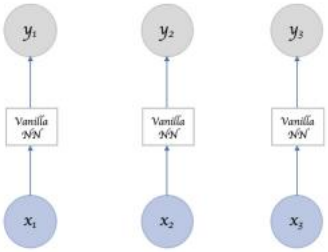
Q14. What are Recurrent Neural Networks (RNNs)?

RNNs are a type of artificial neural networks designed to recognize the pattern from the sequence of data
such as Time series, stock market and government agencies etc.
Recurrent Neural Networks (RNNs) add an interesting twist to basic neural networks. A vanilla neural

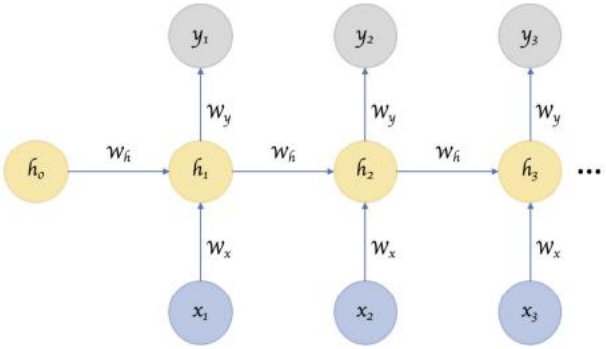
network takes in a fixed size vector as input which limits its usage in situations that involve a ‘series’ type
input with no predetermined size.



RNNs are designed to take a series of input with no predetermined limit on size. One could ask what’s the big deal, I can call a regular NN repeatedly too?



Sure can, but the ‘series’ part of the input means something. A single input item from the series is related
to others and likely has an influence on its neighbors. Otherwise it's just “many” inputs, not a “series” input (duh!).
Recurrent Neural Network remembers the past and its decisions are influenced by what it has learnt from
the past. Note: Basic feed forward networks “remember” things too, but they remember things they learnt during training. For example, an image classifier learns what a “1” looks like during training and then uses that knowledge to classify things in production.
While RNNs learn similarly while training, in addition, they remember things learnt from prior input(s) while generating output(s). RNNs can take one or more input vectors and produce one or more output vectors and the output(s) are influenced not just by weights applied on inputs like a regular NN, but also by a “hidden” state vector representing the context based on prior input(s)/output(s). So, the same input
could produce a different output depending on previous inputs in the series.



In summary, in a vanilla neural network, a fixed size input vector is transformed into a fixed size output vector. Such a network becomes "recurrent" when you repeatedly apply the transformations to a series of given input and produce a series of output vectors. There is no pre-set limitation to the size of the vector. And, in addition to generating the output which is a function of the input and hidden state, we update the hidden state itself based on the input and use it in processing the next input.

Parameter Sharing

You might have noticed another key difference between Figure 1 and Figure 3. In the earlier, multiple different weights are applied to the different parts of an input item generating a hidden layer neuron, which in turn is transformed using further weights to produce an output. There seems to be a lot of weights in play here. Whereas in Figure 3, we seem to be applying the same weights over and over again to different items in the input series.

I am sure you are quick to point out that we are kind of comparing apples and oranges here. The first figure deals with "a" single input whereas the second figure represents multiple inputs from a series. But nevertheless, intuitively speaking, as the number of inputs increase, shouldn't the number of weights in play increase as well? Are we losing some versatility and depth in Figure 3?

Perhaps we are. We are sharing parameters across inputs in Figure 3. If we don't share parameters across
inputs, then it becomes like a vanilla neural network where each input node requires weights of their own.

This introduces the constraint that the length of the input has to be fixed and that makes it impossible to
leverage a series type input where the lengths differ and is not always known.

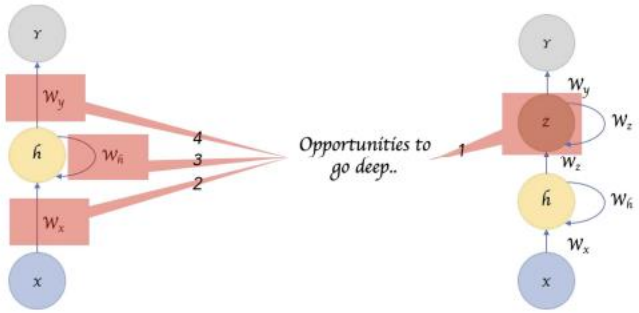
But what we seemingly lose in value here, we gain back by introducing the "hidden state" that links one input to the next. The hidden state captures the relationship that neighbors might have with each other in a serial input and it keeps changing in every step, and thus effectively every input undergoes a different
transition!

Image classifying CNNs have become so successful because the 2D convolutions are an effective form of parameter sharing where each convolutional filter basically extracts the presence or absence of a feature
in an image which is a function of not just one pixel but also of its surrounding neighbor pixels.

In other words, the success of CNNs and RNNs can be attributed to the concept of "parameter sharing" which is fundamentally an effective way of leveraging the relationship between one input item and its surrounding neighbors in a more intrinsic fashion compared to a vanilla neural network.

Deep RNNs

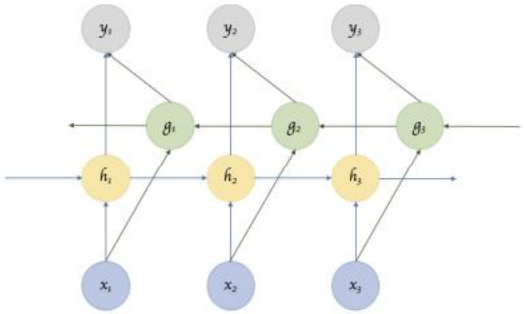
While it's good that the introduction of hidden state enabled us to effectively identify the relationship between the inputs, is there a way we can make an RNN "deep" and gain the multi-level abstractions and
representations we gain through "depth" in a typical neural network?



Here are four possible ways to add depth.
1) We can add hidden states, one on top of another, feeding the output of one to the next.
2) We can also add additional nonlinear hidden layers between input to hidden state.
3) We can increase depth in the hidden to hidden transition.
4) We can increase depth in the hidden to output transition.

Bidirectional RNNs

Sometimes it's not just about learning from the past to predict the future, but we also need to look into the future to fix the past. In speech recognition and handwriting recognition tasks, where there could be considerable ambiguity given just one part of the input, we often need to know what's coming next to better understand the context and detect the present.

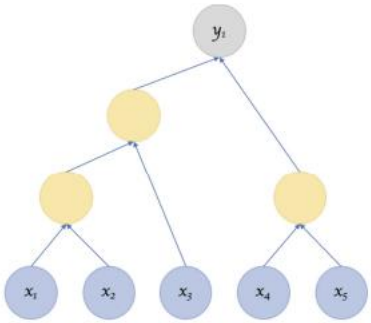


This does introduce the obvious challenge of how much into the future we need to look into, because if we have to wait to see all inputs then the entire operation will become costly. And in cases like speech recognition, waiting till an entire sentence is spoken might make for a less compelling use case. Whereas for NLP tasks, where the inputs tend to be available, we can likely consider entire sentences all at once. Also, depending on the application, if the sensitivity to immediate and closer neighbors is higher than inputs that come further away, a variant that looks only into a limited future/past can be modeled.

Recursive Neural Network

A recurrent neural network parses the inputs in a sequential fashion. A recursive neural network is similar

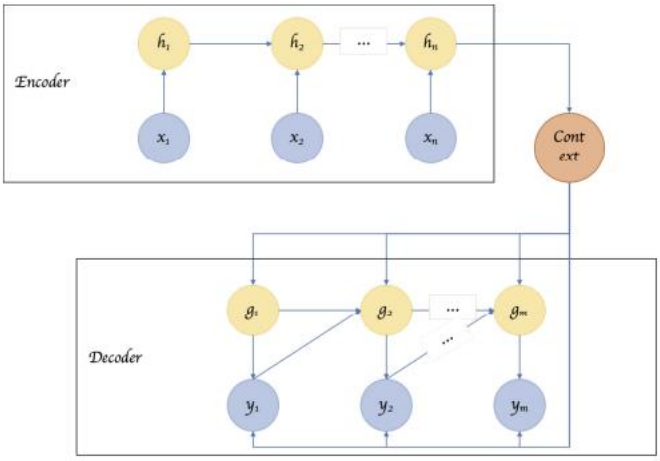
to the extent that the transitions are repeatedly applied to inputs, but not necessarily in a sequential fashion. Recursive Neural Networks are a more general form of Recurrent Neural Networks. It can operate
on any hierarchical tree structure. Parsing through input nodes, combining child nodes into parent nodes
and combining them with other child/parent nodes to create a tree like structure. Recurrent Neural Networks do the same, but the structure there is strictly linear. i.e. weights are applied on the first input node, then the second, third and so on.



But this raises questions pertaining to the structure. How do we decide that? If the structure is fixed like in Recurrent Neural Networks then the process of training, backprop, makes sense in that they are similar
to a regular neural network. But if the structure isn't fixed, is that learnt as well?

Encoder Decoder Sequence to Sequence RNNs

Encoder Decoder or Sequence to Sequence RNNs are used a lot in translation services. The basic idea is that there are two RNNs, one an encoder that keeps updating its hidden state and produces a final single
"Context" output. This is then fed to the decoder, which translates this context to a sequence of outputs.
Another key difference in this arrangement is that the length of the input sequence and the length of the
output sequence need not necessarily be the same.



LSTMs

LSTM is not a different variant of RNN architecture, but rather it introduces changes to how we compute outputs and hidden state using the inputs.

In a vanilla RNN, the input and the hidden state are simply passed through a single tanh layer. LSTM (Long-
Short-Term Memory) networks improve on this simple transformation and introduces additional gates and a cell state, such that it fundamentally addresses the problem of keeping or resetting context, across
sentences and regardless of the distance between such context resets. There are variants of LSTMs including GRUs that utilize the gates in different manners to address the problem of long-term dependencies.

Q15. How Does an LSTM Network Work?

Long-Short-Term Memory (LSTM) is a special kind of recurrent neural network capable of learning longterm
dependencies, remembering information for long periods as its default behavior. There are three steps in an LSTM network:

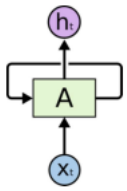
- Step 1: The network decides what to forget and what to remember.
- Step 2: It selectively updates cell state values.
- Step 3: The network decides what part of the current state makes it to the output.

Recurrent Neural Networks

Humans don't start their thinking from scratch every second. As you read this essay, you understand each
word based on your understanding of previous words. You don't throw everything away and start thinking
from scratch again. Your thoughts have persistence.

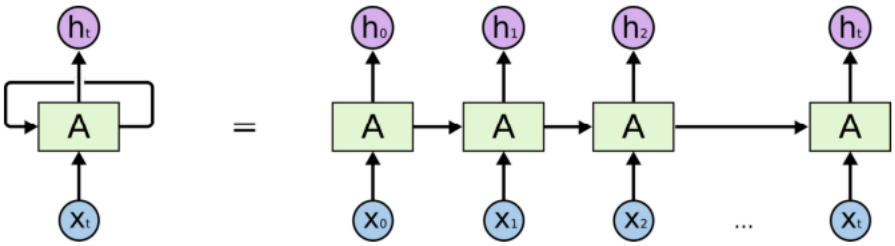
Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional
neural network could use its reasoning about previous events in the film to inform later ones.

Recurrent neural networks address this issue. They are networks with loops in them, allowing information
to persist.



Recurrent Neural Networks have loops.

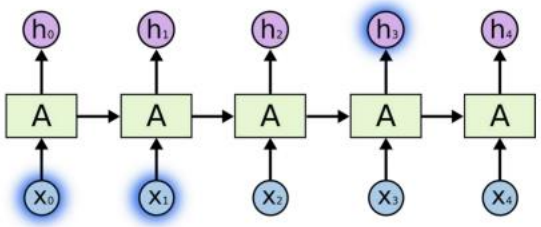
In the above diagram, a chunk of neural network, A, looks at some input □□ and outputs a value h□□. A loop:



allows information to be passed from one step of the network to the next.
These loops make recurrent neural networks seem kind of mysterious. However, if you think a bit more, it turns out that they aren't all that different than a normal neural network. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor. Consider
what happens if we unroll the loop:
An unrolled recurrent neural network.
This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists.
They're the natural architecture of neural network to use for such data.
And they certainly are used! In the last few years, there have been incredible success applying RNNs to a variety of problems: speech recognition, language modeling, translation, image captioning…
Essential to these successes is the use of "LSTMs," a very special kind of recurrent neural network which works, for many tasks, much better than the standard version. Almost all exciting results based on recurrent neural networks are achieved with them.

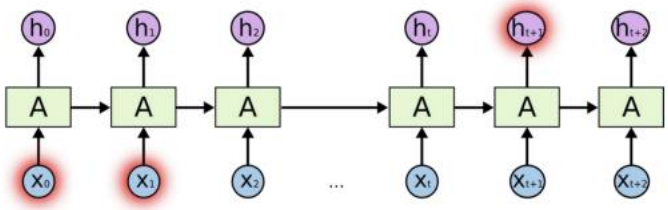
The Problem of Long-Term Dependencies

One of the appeals of RNNs is the idea that they might be able to connect previous information to the present task, such as using previous video frames might inform the understanding of the present frame. If RNNs could do this, they'd be extremely useful. But can they? It depends.
Sometimes, we only need to look at recent information to perform the present task. For example, consider
a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in "the clouds are in the sky," we don't need any further context – it's pretty obvious the next
word is going to be sky. In such cases, where the gap between the relevant information and the place that
it's needed is small, RNNs can learn to use the past information.



But there are also cases where we need more context. Consider trying to predict the last word in the text

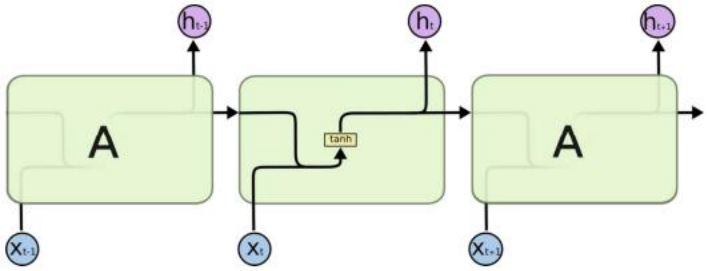
“I grew up in France… I speak fluent French.” Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back. It’s entirely possible for the gap between the relevant information and the point where
it is needed to become very large.
Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.



In theory, RNNs are absolutely capable of handling such “long-term dependencies.” A human could carefully pick parameters for them to solve toy problems of this form. Sadly, in practice, RNNs don’t seem
to be able to learn them. Thankfully, LSTMs don’t have this problem!

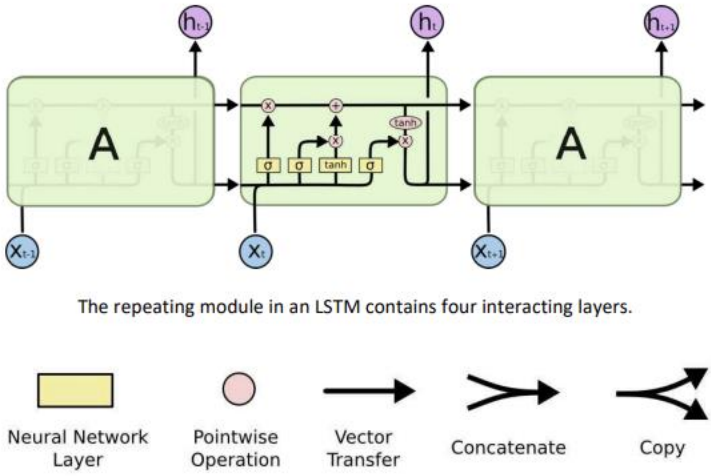
LSTM Networks

Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used.
LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for
long periods of time is practically their default behavior, not something they struggle to learn!
All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.



The repeating module in a standard RNN contains a single layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.
The repeating module in an LSTM contains four interacting layers.



The repeating module in an LSTM contains four interacting layers.

In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denotes its content being copied and the copies going to different locations.

The Core Idea Behind LSTMs

The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.

The cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It's very easy for information to just flow along it unchanged.

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation.

The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means "let nothing through," while a value of one means "let everything through!"

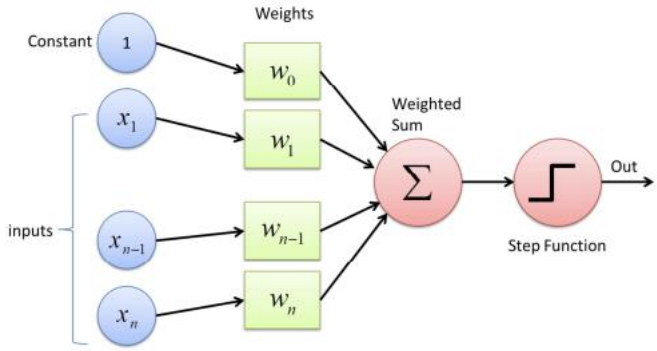
An LSTM has three of these gates, to protect and control the cell state.

Q16. What Is a Multi-layer Perceptron (MLP)?

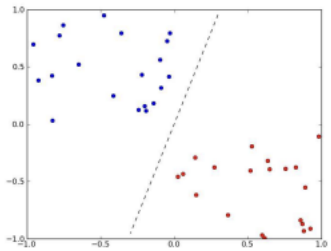
As in Neural Networks, MLPs have an input layer, a hidden layer, and an output layer. It has the same structure as a single layer perceptron with one or more hidden layers.

Perceptron is a single layer neural network and a multi-layer perceptron is called Neural Networks.

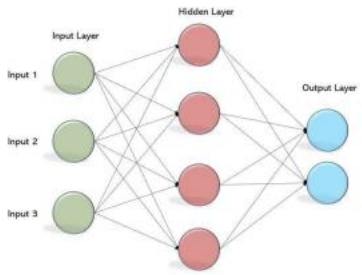
A (single layer) perceptron is a single layer neural network that works as a linear binary classifier. Being a single layer neural network, it can be trained without the use of more advanced algorithms like back propagation and instead can be trained by "stepping towards" your error in steps specified by a learning rate. When someone says perceptron, I usually think of the single layer version.



A single layer perceptron can classify only linear separable classes with binary output {0,1} or {-1,1}, but MLP can classify nonlinear classes. The activation functions are used to map the input between the required values like {0, 1} or {-1, 1}.



Except for the input layer, each node in the other layers uses a nonlinear activation function. This means the input layers, the data coming in, and the activation function is based upon all nodes and weights being
added together, producing the output. MLP uses a supervised learning method called "backpropagation."
In backpropagation, the neural network calculates the error with the help of cost function. It propagates this error backward from where it came (adjusts the weights to train the model more accurately).

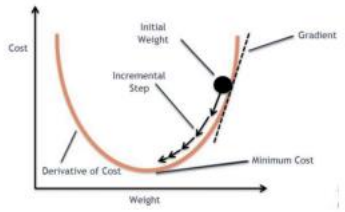


Usually, RELU is in hidden layers (it does not classify), and Soft-max or tanh is in output layers.

Q17. Explain Gradient Descent.

Let's first explain what a gradient is. A gradient is a mathematical function. When calculated on a point of
a function, it gives the hyperplane (or slope) of the directions in which the function increases more. The gradient vector can be interpreted as the "direction and rate of fastest increase". If the gradient of a

function is non-zero at a point p, the direction of the gradient is the direction in which the function increases most quickly from p, and the magnitude of the gradient is the rate of increase in that direction.
Further, the gradient is the zero vector at a point if and only if it is a stationary point (where the derivative
vanishes).
In DS, it simply measures the change in all weights with regard to the change in error, as we are partially derivating by w the loss function.



Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. The goal of the gradient descent is to minimize a given function which, in our case, is the loss function of the neural network. To achieve this goal, it performs two steps iteratively.
1. Compute the slope (gradient) that is the first-order derivative of the function at the current point
2. Move-in the opposite direction of the slope increase from the current point by the computed amount
So, the idea is to pass the training set through the hidden layers of the neural network and then update the parameters of the layers by computing the gradients using the training samples from the training dataset.
Think of it like this. Suppose a man is at top of the valley and he wants to get to the bottom of the valley. So, he goes down the slope. He decides his next position based on his current position and stops when he
gets to the bottom of the valley which was his goal.

Q18. What is exploding gradients?

While training an RNN, if you see exponentially growing (very large) error gradients which accumulate and
result in very large updates to neural network model weights during training, they're known as exploding
gradients. At an extreme, the values of weights can become so large as to overflow and result in NaN values. The explosion occurs through exponential growth by repeatedly multiplying gradients through the
network layers that have values larger than 1.0.
This has the effect of your model is unstable and unable to learn from your training data.
There are some subtle signs that you may be suffering from exploding gradients during the training of your network, such as:
• The model is unable to get traction on your training data (e.g. poor loss).
• The model is unstable, resulting in large changes in loss from update to update.
• The model loss goes to NaN during training.
• The model weights quickly become very large during training.
• The error gradient values are consistently above 1.0 for each node and layer during training.

Solutions

1. Re-Design the Network Model:

a. In deep neural networks, exploding gradients may be addressed by redesigning the network to have fewer layers. There may also be some benefit in using a smaller batch size while training the network.
b. In RNNs, updating across fewer prior time steps during training, called truncated Backpropagation through time, may reduce the exploding gradient problem.
2. Use Long Short-Term Memory Networks: In RNNs, exploding gradients can be reduced by using the Long Short-Term Memory (LSTM) memory units and perhaps related gated-type neuron structures. Adopting LSTM memory units is a new best practice for recurrent neural networks for sequence prediction.
3. Use Gradient Clipping: Exploding gradients can still occur in very deep Multilayer Perceptron networks with a large batch size and LSTMs with very long input sequence lengths. If exploding gradients are still occurring, you can check for and limit the size of gradients during the training of your network. This is called gradient clipping. Specifically, the values of the error gradient are checked against a threshold value and clipped or set to that threshold value if the error gradient exceeds the threshold.
4. Use Weight Regularization: another approach, if exploding gradients are still occurring, is to check the size of network weights and apply a penalty to the networks loss function for large weight values. This is called weight regularization and often an L1 (absolute weights) or an L2 (squared weights) penalty can be used.

Q19. What is vanishing gradients?

While training an RNN, your slope can become either too small; this makes the training difficult. When the slope is too small, the problem is known as a Vanishing Gradient. It leads to long training times, poor performance, and low accuracy.
• Hyperbolic tangent and Sigmoid/Soft-max suffer vanishing gradient.
• RNNs suffer vanishing gradient, LSTM no (so it is perfect to predict stock prices). In fact, the propagation of error through previous layers makes the gradient get smaller so the weights are not updated.

Solutions

1. Choose RELU
2. Use LSTM (for RNNs)
3. Use ResNet (Residual Network) à after some layers, add x again: $F(x) \rightarrow \cdots \rightarrow F(x) + x$
4. Multi-level hierarchy: pre-train one layer at the time through unsupervised learning, then finetune via backpropagation
5. Gradient checking: debugging strategy used to numerically track and assess gradients during training.

Q20. What is Back Propagation and Explain it Works.

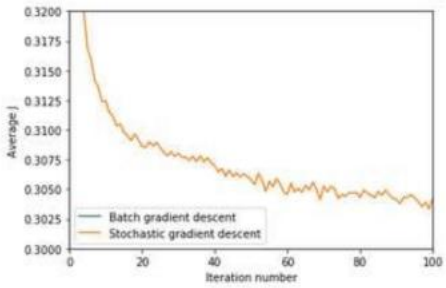
Backpropagation is a training algorithm used for neural network. In this method, we update the weights of each layer from the last layer recursively, with the formula:

$$w_{previous\ layer} = w_{layer} - \eta \nabla_w L(w)$$

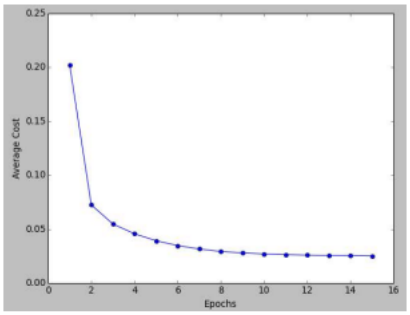
It has the following steps:
• Forward Propagation of Training Data (initializing weights with random or pre-assigned values)
• Gradients are computed using output weights and target
• Back Propagate for computing gradients of error from output activation
• Update the Weights

Q21. What are the variants of Back Propagation?

• Stochastic Gradient Descent: In Batch Gradient Descent we were considering all the examples for every step of Gradient Descent. But what if our dataset is very huge. Deep learning models crave for data. The more the data the more chances of a model to be good. Suppose our dataset has 5 million examples, then just to take one step the model will have to calculate the gradients of all the 5 million examples. This does not seem an efficient way. To tackle this problem, we have Stochastic Gradient Descent. In Stochastic Gradient Descent (SGD), we consider just one example at a time to take a single step. We do the following steps in one epoch for SGD:
1. Take an example
2. Feed it to Neural Network
3. Calculate its gradient
4. Use the gradient we calculated in step 3 to update the weights
5. Repeat steps 1–4 for all the examples in training dataset
Since we are considering just one example at a time the cost will fluctuate over the training examples and it will not necessarily decrease. But in the long run, you will see the cost decreasing with fluctuations. Also, because the cost is so fluctuating, it will never reach the minimum, but it will keep dancing around it. SGD can be used for larger datasets. It converges faster when the dataset is large as it causes updates to the parameters more frequently.



• Batch Gradient Descent: all the training data is taken into consideration to take a single step. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. So that's just one step of gradient descent in one epoch. Batch Gradient Descent is great for convex or relatively smooth error manifolds. In this case, we move somewhat directly towards an optimum solution. The graph of cost vs epochs is also quite smooth because we are averaging over all the gradients of training data for a single step. The cost keeps on decreasing over the epochs.



• Mini-batch Gradient Descent: It's one of the most popular optimization algorithms. It's a variant of Stochastic Gradient Descent and here instead of single training example, mini batch of samples is

used. Batch Gradient Descent can be used for smoother curves. SGD can be used when the dataset is large. Batch Gradient Descent converges directly to minima. SGD converges faster for larger datasets. But, since in SGD we use only one example at a time, we cannot implement the vectorized implementation on it. This can slow down the computations. To tackle this problem, a mixture of Batch Gradient Descent and SGD is used. Neither we use all the dataset all at once nor we use the single example at a time. We use a batch of a fixed number of training examples which is less than the actual dataset and call it a mini-batch. Doing this helps us achieve the advantages of both the former variants we saw. So, after creating the mini-batches of fixed size, we do the following steps in one epoch:

1. Pick a mini-batch
2. Feed it to Neural Network
3. Calculate the mean gradient of the mini-batch
4. Use the mean gradient we calculated in step 3 to update the weights
5. Repeat steps 1–4 for the mini-batches we created

Just like SGD, the average cost over the epochs in mini-batch gradient descent fluctuates because we are averaging a small number of examples at a time. So, when we are using the mini-batch gradient descent we are updating our parameters frequently as well as we can use vectorized implementation for faster computations.

Q22. What are the different Deep Learning Frameworks?

• PyTorch: PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab. It is free and open-source software released under the Modified BSD license.

• TensorFlow: TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library and is also used for machine learning applications such as neural networks. Licensed by Apache License 2.0. Developed by Google Brain Team.

• Microsoft Cognitive Toolkit: Microsoft Cognitive Toolkit describes neural networks as a series of computational steps via a directed graph.

• Keras: Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. Licensed by MIT.

Q23. What is the role of the Activation Function?

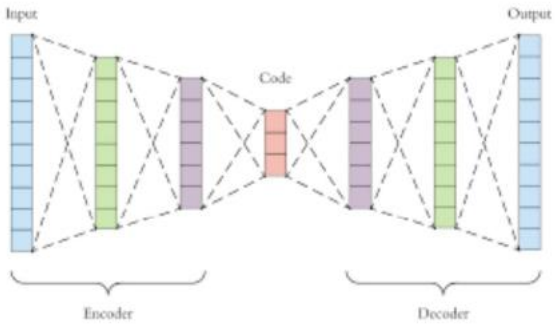
The Activation function is used to introduce non-linearity into the neural network helping it to learn more
complex function. Without which the neural network would be only able to learn linear function which is
a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs.

Q24. Name a few Machine Learning libraries for various purposes.

Purpose	Libraries
Scientific Computation	Numpy
Tabular Data	Pandas, GeoPandas
Data Modelling & Preprocessing	Scikit Learn
Time-Series Analysis	Statsmodels
Text processing	NTLK, Regular Expressions
Deep Learning	TensorFlow, Pytorch
Visualization	Bokeh, Seaborn
Plotting	Matplot

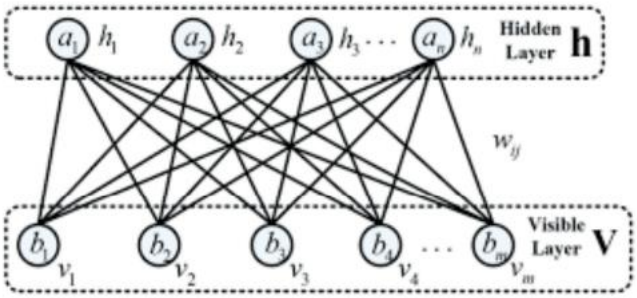
Q25. What is an Auto-Encoder?

Auto-encoders are simple learning networks that aim to transform inputs into outputs with the minimum
possible error. This means that we want the output to be as close to input as possible. We add a couple of layers between the input and the output, and the sizes of these layers are smaller than the input layer.
The auto-encoder receives unlabeled input which is then encoded to reconstruct the input.
An autoencoder is a type of artificial neural network used to learn efficient data coding in an unsupervised
manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”. Along with the reduction side,
a reconstructing side is learnt, where the autoencoder tries to generate from the reduced encoding a representation as close as possible to its original input, hence its name. Several variants exist to the basic
model, with the aim of forcing the learned representations of the input to assume useful properties.
Autoencoders are effectively used for solving many applied problems, from face recognition to acquiring the semantic meaning of words.



Q26. What is a Boltzmann Machine?

Boltzmann machines have a simple learning algorithm that allows them to discover interesting features that represent complex regularities in the training data. The Boltzmann machine is basically used to optimize the weights and the quantity for the given problem. The learning algorithm is very slow in networks with many layers of feature detectors. “Restricted Boltzmann Machines” algorithm has a single
layer of feature detectors which makes it faster than the rest.



Q27. What Is Dropout and Batch Normalization?

Dropout is a technique of dropping out hidden and visible nodes of a network randomly to prevent overfitting of data (typically dropping 20 per cent of the nodes). It doubles the number of iterations needed to converge the network. It used to avoid overfitting, as it increases the capacity of generalization.

Batch normalization is the technique to improve the performance and stability of neural networks by normalizing the inputs in every layer so that they have mean output activation of zero and standard deviation of one.

Q28. Why Is TensorFlow the Most Preferred Library in Deep Learning?

TensorFlow provides both C++ and Python APIs, making it easier to work on and has a faster compilation time compared to other Deep Learning libraries like Keras and PyTorch. TensorFlow supports both CPU and GPU computing devices.

Q29. What Do You Mean by Tensor in TensorFlow?

A tensor is a mathematical object represented as arrays of higher dimensions. Think of a n-D matrix. These
arrays of data with different dimensions and ranks fed as input to the neural network are called “Tensors.”

Q30. What is the Computational Graph?

Everything in a TensorFlow is based on creating a computational graph. It has a network of nodes where each node operates. Nodes represent mathematical operations, and edges represent tensors. Since data flows in the form of a graph, it is also called a “DataFlow Graph.”

Q31. How is logistic regression done?

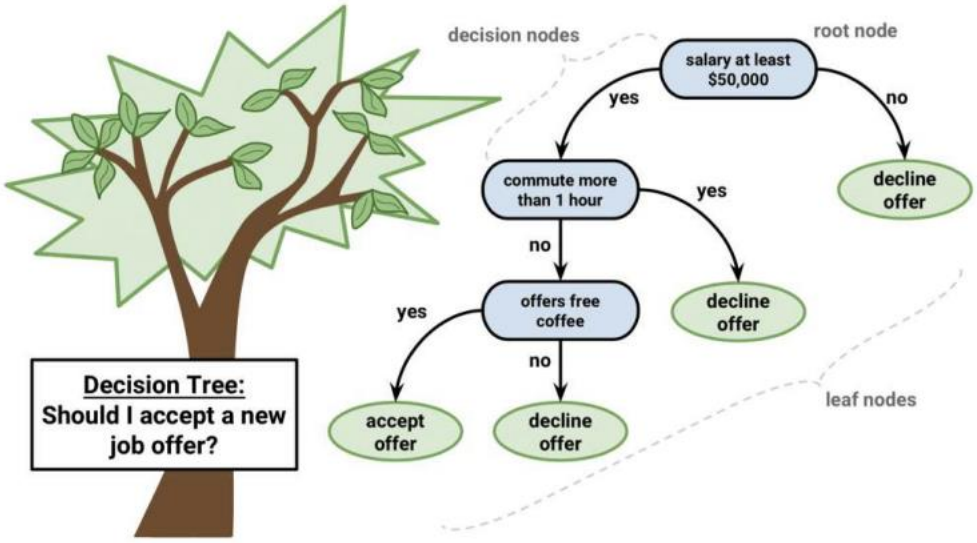
Logistic regression measures the relationship between the dependent variable (our label of what we want
to predict) and one or more independent variables (our features) by estimating probability using its underlying logistic function (sigmoid).

Miscellaneous

Q1. Explain the steps in making a decision tree.

1. Take the entire data set as input
2. Calculate entropy of the target variable, as well as the predictor attributes
3. Calculate your information gain of all attributes (we gain information on sorting different objects from each other)
4. Choose the attribute with the highest information gain as the root node
5. Repeat the same procedure on every branch until the decision node of each branch is finalized

For example, let's say you want to build a decision tree to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:



It is clear from the decision tree that an offer is accepted if:

- Salary is greater than $50,000
- The commute is less than an hour
- Coffee is offered

Q2. How do you build a random forest model?

A random forest is built up of a number of decision trees. If you split the data into different packages and
make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

1. Randomly select □ features from a total of □ features where □ << □
2. Among the □ features, calculate the node D using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat steps two and three until leaf nodes are finalized
5. Build forest by repeating steps one to four for □ times to create □ number of trees

Q3. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Example: height of students

Height (in cm)
164
167.3
170
174.2
178
180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celsius)	Sales (in K $)
20	2.0
25	2.1
26	2.3
28	2.7
30	3.1

Here, the relationship is visible from the table that temperature and sales are directly proportional to each
other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.
Example: data for house price prediction
The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range,
minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

Q4. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection.

Filter Methods

This involves:

- Linear discrimination analysis
- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting
the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit
- Backward Selection: We test all the features and start removing them to see what works better
- Recursive Feature Elimination: Recursively looks through all the different features and how they pair together

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is
performed with the wrapper method.

Q5. In your choice of language, write a program that prints the numbers ranging from one to 50. But for multiples of three, print "Fizz" instead of the number and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz."

The code is shown below:

```
for x in range(51):

      if x % 3 == 0 and x % 5 == 0:
            print('fizzbuzz')

      elif x % 3 == 0:
            print('fizz')

      elif x % 5 == 0:
            print('buzz')

      else:
            print('fizzbuzz')
```

Q6. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way;
we use the rest of the data to predict the values.
For smaller data sets, we can impute missing values with the mean, median, or average of the rest of the data using pandas data frame in python. There are different ways to do so, such as:
df.mean(), df.fillna(mean)
Other option of imputation is using KNN for numeric or classification values (as KNN just uses k closest values to impute the missing value).

Q7. For the given points, how will you calculate the Euclidean distance in Python?

plot1 = [1,3]
plot2 = [2,5]
The Euclidean distance can be calculated as follows:
euclidean_distance = sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)

Q8. What are dimensionality reduction and its benefits?

Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.
This reduction helps in compressing data and reducing storage space. It also reduces computation time as
fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in
storing a value in two different units (meters and inches).

Q9. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

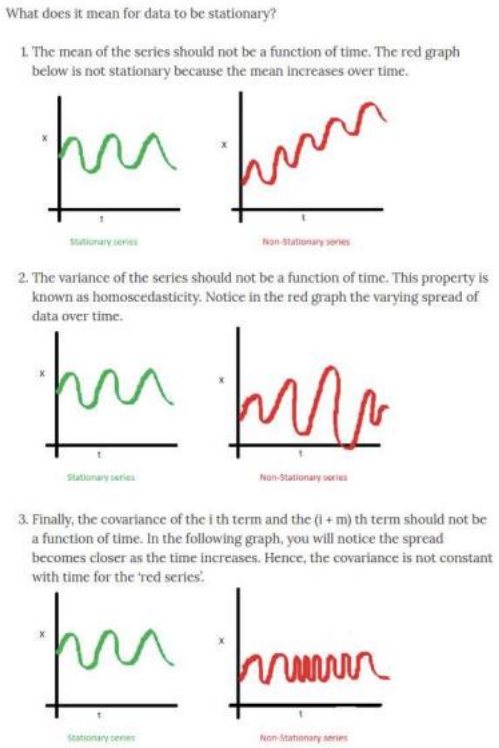
Determinant of A − $\lambda 1$ and solve to find λ .

Q10. How should you maintain a deployed model?

The steps to maintain a deployed model are (CREM):
1. Monitor: constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.
2. Evaluate: evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

3. Compare: the new models are compared to each other to determine which model performs the best.
4. Rebuild: the best performing model is re-built on the current state of data.

Q11. How can a time-series data be declared as stationery?



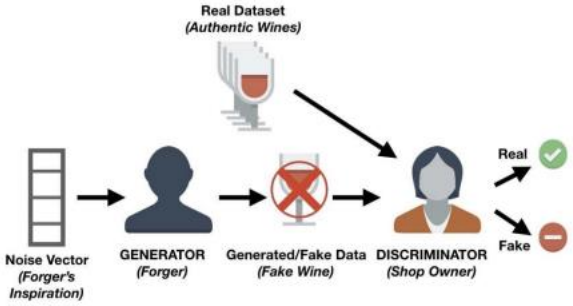
Q12. 'People who bought this also bought...' recommendations seen on Amazon are a result of which algorithm?

The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.
The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.
For example, a sales page shows that a certain number of people buy a new phone and also buy tempered
glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to
buy tempered glass as well.

Q13. What is a Generative Adversarial Network?

Suppose there is a wine shop purchasing wine from dealers, which they resell later. But some dealers sell
fake wine. In this case, the shop owner should be able to distinguish between fake and authentic wine. The
forger will try different techniques to sell fake wine and make sure specific techniques go past the shop owner’s check. The shop owner would probably get some feedback from wine experts that some of the wine is not original. The owner would have to improve how he determines whether a wine is fake or authentic.

The forger's goal is to create wines that are indistinguishable from the authentic ones while the shop owner
intends to tell if the wine is real or not accurately.



• There is a noise vector coming into the forger who is generating fake wine.
• Here the forger acts as a Generator.
• The shop owner acts as a Discriminator.
• The Discriminator gets two inputs; one is the fake wine, while the other is the real authentic wine.
The shop owner has to figure out whether it is real or fake.
So, there are two primary components of Generative Adversarial Network (GAN) named:
1. Generator
2. Discriminator
The generator is a CNN that keeps keys producing images and is closer in appearance to the real images while the discriminator tries to determine the difference between real and fake images. The ultimate aim
is to make the discriminator learn to identify real and fake images.

Q14. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a
measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection and can greatly improve a patient's prognosis.
Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

Q15. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable? [0, 0, 0, 1, 1, 1, 1, 1]

The target variable, in this case, is 1 (the last)

The formula for calculating the entropy is, putting $p = 5$ and $n = 8$, we get:

$$Entropy = -\left(\frac{5}{8}\, log\left(\frac{5}{8}\right) + \frac{3}{8}\log\left(\frac{3}{8}\right)\right)$$

Q16. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case? Choose the correct option:

The most appropriate algorithm for this case is logistic regression.

Q17. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?

As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering is the most appropriate algorithm for this study.

Q18. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true? Choose the right answer:

The answer is A: {grape, apple} must be a frequent itemset.

Q19. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

One-way ANOVA: in statistics, one-way analysis of variance is a technique that can be used to compare means of two or more samples. This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or categorical input data, the "X", always one variable, hence "oneway".

The ANOVA tests the null hypothesis, which states that samples in all groups are drawn from populations

with the same mean values. To do this, two estimates are made of the population variance. The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values.

Q20. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an

object in a mathematical way that's easy to analyze.

Q21. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other

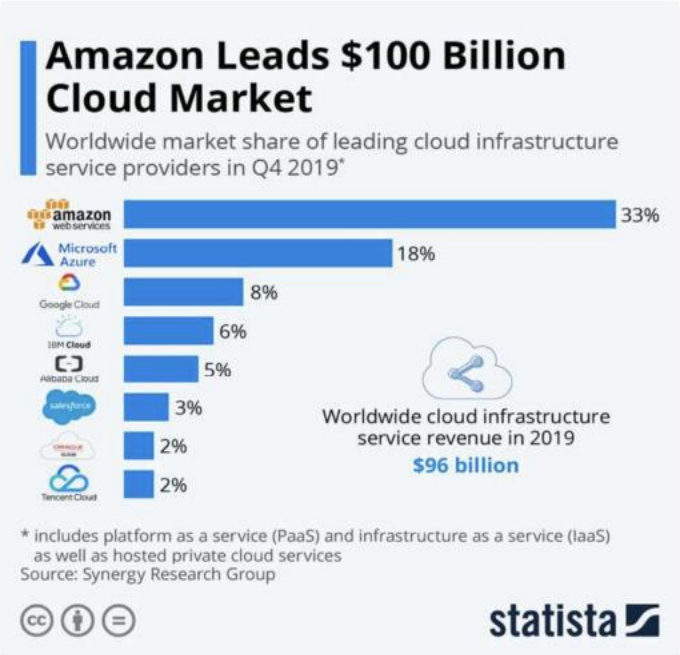
areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

Q22. Do gradient descent methods always converge to similar points?

They do not, because in some cases, they reach a local minimum or a local optimum point. You would not

reach the global optimum point. This is governed by the data and the starting conditions.

Q23. What are the most popular Cloud Services used in Data Science?



Q24. What is a Canary Deployment?

A canary deployment, or canary release, allows you to rollout your features to only a subset of users as an initial test to make sure nothing else in your system broke.
The initial steps for implementing canary deployment are:
1. create two clones of the production environment,
2. have a load balancer that initially sends all traffic to one version,
3. create new functionality in the other version.
When you deploy the new software version, you shift some percentage – say, 10% – of your user base to the new version while maintaining 90% of users on the old version. If that 10% reports no errors, you can
roll it out to gradually more users, until the new version is being used by everyone. If the 10% has problems, though, you can roll it right back, and 90% of your users will have never even seen the problem.
Canary deployment benefits include zero downtime, easy rollout and quick rollback – plus the added safety from the gradual rollout process. It also has some drawbacks – the expense of maintaining multiple
server instances, the difficult clone-or-don't-clone database decision.
Typically, software development teams implement blue/green deployment when they're sure the new version will work properly and want a simple, fast strategy to deploy it. Conversely, canary deployment is
most useful when the development team isn't as sure about the new version and they don't mind a slower
rollout if it means they'll be able to catch the bugs.

Q25. What is a Blue Green Deployment?

Blue-green deployment is a technique that reduces downtime and risk by running two identical production environments called Blue and Green.

At any time, only one of the environments is live, with the live environment serving all production traffic.
For this example, Blue is currently live, and Green is idle.
As you prepare a new version of your model, deployment and the final stage of testing takes place in the environment that is not live: in this example, Green. Once you have deployed and fully tested the model in Green, you switch the router, so all incoming requests now go to Green instead of Blue. Green is now live, and Blue is idle.
This technique can eliminate downtime due to app deployment and reduces risk: if something unexpected
happens with your new version on Green, you can immediately roll back to the last version by switching back to Blue.

..

Interview Question Series #2
Python Programming

Numpy

1. Why is python numpy better than lists?

Python numpy arrays should be considered instead of a list because they are fast, consume less memory and convenient with lots of functionality.

2. Describe the map function in Python?

map function executes the function given as the first argument on all the elements of the iterable given as the second argument.

3. Generate array of '100' random numbers sampled from a standard normal distribution using Numpy

np.random.rand(100) will create 100 random numbers generated from standard normal distribution with mean 0 and standard deviation 1.

4. How to count the occurrence of each value in a numpy array?

Use numpy.bincount()

```
>>> arr = numpy.array([0, 5, 5, 0, 2, 4, 3, 0, 0, 5, 4, 1, 9, 9])
>>> numpy.bincount(arr)
```

The argument to bincount() must consist of booleans or positive integers. Negative integers are invalid.

5. Does Numpy Support Nan?

nan, short for "not a number", is a special floating point value defined by the IEEE-754 specification. Python numpy supports nan but the definition of nan is more system dependent and some systems don't have an all round support for it like older cray and vax computers.

6. What does ravel() function in numpy do?

It combines multiple numpy arrays into a single array

7. What is the meaning of axis=0 and axis=1?

Axis = 0 is meant for reading rows, Axis = 1 is meant for reading columns

..

8. What is numpy and describe its use cases?

Numpy is a package library for Python, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high level mathematical functions. In simple

words, Numpy is an optimized version of Python lists like Financial functions, Linear Algebra, Statistics, Polynomials, Sorting and Searching etc.

9. How to remove from one array those items that exist in another?

```
>>> a = np.array([5, 4, 3, 2, 1])
>>> b = np.array([4, 8, 9, 10, 1])
# From 'a' remove all of 'b'
>>> np.setdiff1d(a,b)
# Output:
>>> array([5, 3, 2])
```

10. How to sort a numpy array by a specific column in a 2D array?

```
#Choose column 2 as an example
>>> import numpy as np
>>> arr = np.array([[1, 2, 3], [4, 5, 6], [0,0,1]])
>>> arr[arr[:,1].argsort()]
# Output
>>> array([[0, 0, 1], [1, 2, 3], [4, 5, 6]])
```

11. How to reverse a numpy array in the most efficient way?

```
>>> import numpy as np
>>> arr = np.array([9, 10, 1, 2, 0])
>>> reverse_arr = arr[::-1]
```

12. How to calculate percentiles when using numpy?

```
>>> import numpy as np
>>> arr = np.array([11, 22, 33, 44 ,55 ,66, 77])
>>> perc = np.percentile(arr, 40) #Returns the 40th percentile
>>> print(perc)
```

13.What Is The Difference Between Numpy And Scipy?

NumPy would contain nothing but the array data type and the most basic operations: indexing, sorting, reshaping, basic element wise functions, et cetera. All numerical code would reside in SciPy. SciPy contains more fully-featured versions of the linear algebra modules, as well as many other numerical algorithms.

..

14. What Is The Preferred Way To Check For An Empty (zero Element) Array?

For a numpy array, use the size attribute. The size attribute is helpful for determining the length of numpy array:

```
>>> arr = numpy.zeros((1,0))
>>> arr.size
```

15. What Is The Difference Between Matrices And Arrays?

Matrices can only be two-dimensional, whereas arrays can have any number of dimensions

16. How can you find the indices of an array where a condition is true?

Given an array a, the condition arr > 3 returns a boolean array and since False is interpreted as 0 in Python and NumPy.

```
>>> import numpy as np
>>> arr = np.array([[9,8,7],[6,5,4],[3,2,1]])
>>> arr > 3
>>> array([[True, True, True],
[ True, True, True],
```

[False, False, False]], dtype=bool)

17. How to find the maximum and minimum value of a given flattened array?

```
>>> import numpy as np
>>> a = np.arange(4).reshape((2,2))
>>> max_val = np.amax(a)
>>> min_val = np.amin(a)
```

18. Write a NumPy program to calculate the difference between the maximum and the minimum values of a given array along the second axis.

```
>>> import numpy as np
>>> arr = np.arange(16).reshape((4, 7))
>>> res = np.ptp(arr, 1)
```

19. Find median of a numpy flattened array

```
>>> import numpy as np
>>> arr = np.arange(16).reshape((4, 5))
>>> res = np.median(arr)
```

..

20. Write a NumPy program to compute the mean, standard deviation, and variance of a given array along the second axis

```
import numpy as np
>>> import numpy as np
>>> x = np.arange(16)
>>> mean = np.mean(x)
>>> std = np.std(x)
>>> var= np.var(x)
```

21. Calculate covariance matrix between two numpy arrays

```
>>> import numpy as np
>>> x = np.array([2, 1, 0])
>>> y = np.array([2, 3, 3])
>>> cov_arr = np.cov(x, y)
```

22. Compute Compute pearson product-moment correlation coefficients of two given numpy arrays

```
>>> import numpy as np
>>> x = np.array([0, 1, 3])
>>> y = np.array([2, 4, 5])
>>> cross_corr = np.corrcoef(x, y)
```

23. Develop a numpy program to compute the histogram of nums against the bins

```
>>> import numpy as np
>>> nums = np.array([0.5, 0.7, 1.0, 1.2, 1.3, 2.1])
>>> bins = np.array([0, 1, 2, 3])
>>> np.histogram(nums, bins)
```

24. Get the powers of an array values element-wise

```
>>> import numpy as np
>>> x = np.arange(7)
>>> np.power(x, 3)
```

25. Write a NumPy program to get true division of the element-wise array inputs

```
>>> import numpy as np
>>> x = np.arange(10)
```

```
>>> np.true_divide(x, 3)
```
..

Pandas

26.What is a series in pandas?

A Series is defined as a one-dimensional array that is capable of storing various data types. The row labels of the series are called the index. By using a 'series' method, we can easily convert the list, tuple, and dictionary into series. A Series cannot contain multiple columns.

27.What features make Pandas such a reliable option to store tabular data?

Memory Efficient, Data Alignment, Reshaping, Merge and join and Time Series.

28.What is reindexing in pandas?

Reindexing is used to conform DataFrame to a new index with optional filling logic. It places NA/NaN in that location where the values are not present in the previous index. It returns a new object unless the new index is produced as equivalent to the current one, and the value of copy becomes False. It is used to change the index of the rows and columns of the DataFrame.

29. How will you create a series from dict in Pandas?

A Series is defined as a one-dimensional array that is capable of storing various data types.

```
>>> import pandas as pd
>>> info = {'x' : 0., 'y' : 1., 'z' : 2.}
>>> a = pd.Series(info)
```

30. How can we create a copy of the series in Pandas?

Use pandas.Series.copy method

```
>>> import pandas as pd
>>> pd.Series.copy(deep=True)
```

31.What is groupby in Pandas?

GroupBy is used to split the data into groups. It groups the data based on some criteria. Grouping also provides a mapping of labels to the group names. It has a lot of variations that can be defined with the parameters and makes the task of splitting the data quick and easy.

32.What is vectorization in Pandas?

..

Vectorization is the process of running operations on the entire array. This is done to reduce the amount of iteration performed by the functions. Pandas have a number of vectorized functions like aggregations, and string functions that are optimized to operate specifically on series and DataFrames. So it is preferred to use the vectorized pandas functions to execute the operations quickly.

33. Mention the different types of Data Structures in Pandas

Pandas provide two data structures, which are supported by the pandas library, Series, and DataFrames. Both of these data structures are built on top of the NumPy.

34.What Is Time Series In pandas

A time series is an ordered sequence of data which basically represents how some quantity changes over time. pandas contains extensive capabilities and features for working with time series data for all domains.

35. How to convert pandas dataframe to numpy array?

The function to_numpy() is used to convert the DataFrame to a NumPy array.
DataFrame.to_numpy(self, dtype=None, copy=False)
The dtype parameter defines the data type to pass to the array and the copy ensures the returned value is not a view on another array.

36.Write a Pandas program to get the first 5 rows of a given DataFrame

```
>>> import pandas as pd
>>> exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],}
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
>>> df = pd.DataFrame(exam_data , index=labels)
>>> df.iloc[:5]
```

37. Develop a Pandas program to create and display a one-dimensional arraylike object containing an array of data.

```
>>> import pandas as pd
>>> pd.Series([2, 4, 6, 8, 10])
```

38.Write a Python program to convert a Panda module Series to Python list and it's type.

```
>>> import pandas as pd
>>> ds = pd.Series([2, 4, 6, 8, 10])
..
>>> type(ds)
>>> ds.tolist()
>>> type(ds.tolist())
```

39. Develop a Pandas program to add, subtract, multiple and divide two Pandas Series.

```
>>> import pandas as pd
>>> ds1 = pd.Series([2, 4, 6, 8, 10])
>>> ds2 = pd.Series([1, 3, 5, 7, 9])
>>> sum = ds1 + ds2
>>> sub = ds1 - ds2
>>> mul = ds1 * ds2
>>> div = ds1 / ds2
```

40. Develop a Pandas program to compare the elements of the two Pandas Series.

```
>>> import pandas as pd
>>> ds1 = pd.Series([2, 4, 6, 8, 10])
>>> ds2 = pd.Series([1, 3, 5, 7, 10])
>>> ds1 == ds2
>>> ds1 > ds2
>>> ds1 < ds2
```

41. Develop a Pandas program to change the data type of given a column or a Series.

```
>>> import pandas as pd
>>> s1 = pd.Series(['100', '200', 'python', '300.12', '400'])
>>> s2 = pd.to_numeric(s1, errors='coerce')
>>> s2
```

42.Write a Pandas program to convert Series of lists to one Series

```
>>> import pandas as pd
>>> s = pd.Series([ ['Red', 'Black'], ['Red', 'Green', 'White'] , ['Yellow']])
>>> s = s.apply(pd.Series).stack().reset_index(drop=True)
```
43.Write a Pandas program to create a subset of a given series based on value and condition
```
>>> import pandas as pd
>>> s = pd.Series([0, 1,2,3,4,5,6,7,8,9,10])
>>> n = 6
..
>>> new_s = s[s < n]
>>> new_s
```
44. Develop a Pandas code to alter the order of index in a given series
```
>>> import pandas as pd
>>> s = pd.Series(data = [1,2,3,4,5], index = ['A', 'B', 'C','D','E'])
>>> s.reindex(index = ['B','A','C','D','E'])
```
45.Write a Pandas code to get the items of a given series not present in another given series.
```
>>> import pandas as pd
>>> sr1 = pd.Series([1, 2, 3, 4, 5])
>>> sr2 = pd.Series([2, 4, 6, 8, 10])
>>> result = sr1[~sr1.isin(sr2)]
>>> result
```
46.What is the difference between the two data series df['Name'] and df.loc[:, 'Name']?
```
>>> First one is a view of the original dataframe and second one is a copy of the original dataframe.
```
47.Write a Pandas program to display the most frequent value in a given series and replace everything else as "replaced" in the series.
```
>>> import pandas as pd
>>> import numpy as np
>>> np.random.RandomState(100)
>>> num_series = pd.Series(np.random.randint(1, 5, [15]))
>>> result = num_series[~num_series.isin(num_series.value_counts().index[:1])] =
'replaced'
```
48.Write a Pandas program to find the positions of numbers that are multiples of 5 of a given series.
```
>>> import pandas as pd
>>> import numpy as np
>>> num_series = pd.Series(np.random.randint(1, 10, 9))
>>> result = np.argwhere(num_series % 5==0)
```
49. How will you add a column to a pandas DataFrame?
```
..
# importing the pandas library
>>> import pandas as pd
>>> info = {'one' : pd.Series([1, 2, 3, 4, 5], index=['a', 'b', 'c', 'd', 'e']),
'two' : pd.Series([1, 2, 3, 4, 5, 6], index=['a', 'b', 'c', 'd', 'e', 'f'])}
>>> info = pd.DataFrame(info)
```

```
# Add a new column to an existing DataFrame object
>>> info['three']=pd.Series([20,40,60],index=['a','b','c'])
```

50. How to iterate over a Pandas DataFrame?

You can iterate over the rows of the DataFrame by using for loop in combination with an iterrows() call on the DataFrame.

Python Language

51.What type of language is python? Programming or scripting?

Python is capable of scripting, but in general sense, it is considered as a general-purpose programming language.

52. Is python case sensitive?

Yes, python is a case sensitive language.

53.What is a lambda function in python?

An anonymous function is known as a lambda function. This function can have any number of parameters but can have just one statement.

54.What is the difference between xrange and xrange in python?

xrange and range are the exact same in terms of functionality.The only difference is that range returns a Python list object and x range returns an xrange object.

55.What are docstrings in python?

Docstrings are not actually comments, but they are documentation strings. These docstrings are within triple quotes. They are not assigned to any variable and therefore, at times, serve the purpose of comments as well.

56.Whenever Python exits, why isn't all the memory deallocated?

Whenever Python exits, especially those Python modules which are having circular references to other objects or the objects that are referenced from the global namespaces ..
are not always de-allocated or freed. It is impossible to de-allocate those portions of memory that are reserved by the C library. On exit, because of having its own efficient clean up mechanism, Python would try to de-allocate/destroy every other object.

57.What does this mean: *args, **kwargs? And why would we use it?

We use *args when we aren't sure how many arguments are going to be passed to a function, or if we want to pass a stored list or tuple of arguments to a function. **kwargs is used when we don't know how many keyword arguments will be passed to a function, or it can be used to pass the values of a dictionary as keyword arguments.

58.What is the difference between deep and shallow copy?

Shallow copy is used when a new instance type gets created and it keeps the values that are copied in the new instance. Shallow copy is used to copy the reference pointers just like it copies the values.
Deep copy is used to store the values that are already copied. Deep copy doesn't copy the reference pointers to the objects. It makes the reference to an object and the new object that is pointed by some other object gets stored.

59. Define encapsulation in Python?

Encapsulation means binding the code and the data together. A Python class in an example of encapsulation.

60. Does python make use of access specifiers?

Python does not deprive access to an instance variable or function. Python lays down the concept of prefixing the name of the variable, function or method with a single or double

underscore to imitate the behavior of protected and private access specifiers.

61.What are the generators in Python?

Generators are a way of implementing iterators. A generator function is a normal function except that it contains yield expression in the function definition making it a generator function.

62. How will you remove the duplicate elements from the given list?

The set is another type available in Python. It doesn't allow copies and provides some good functions to perform set operations like union, difference etc.

```
>>> list(set(a))
```

63. Does Python allow arguments Pass by Value or Pass by Reference?

..

Neither the arguments are Pass by Value nor does Python supports Pass by reference. Instead, they are Pass by assignment. The parameter which you pass is originally a reference to the object not the reference to a fixed memory location. But the reference is passed by value. Additionally, some data types like strings and tuples are immutable whereas others are mutable.

64.What is slicing in Python?

Slicing in Python is a mechanism to select a range of items from Sequence types like strings, list, tuple, etc.

65.Why is the "pass" keyword used in Python?

The "pass" keyword is a no-operation statement in Python. It signals that no action is required. It works as a placeholder in compound statements which are intentionally left blank.

66.What is PEP8 and why is it important?

PEP stands for Python Enhancement Proposal. A PEP is an official design document providing information to the Python Community, or describing a new feature for Python or its processes. PEP 8 is especially important since it documents the style guidelines for Python Code. Apparently contributing in the Python open-source community requires you to follow these style guidelines sincerely and strictly.

67.What are decorators in Python?

Decorators in Python are essentially functions that add functionality to an existing function in Python without changing the structure of the function itself. They are represented by the @decorator_name in Python and are called in bottom-up fashion

68.What is the key difference between lists and tuples in python?

The key difference between the two is that while lists are mutable, tuples on the other hand are immutable objects.

69.What is self in Python?

Self is a keyword in Python used to define an instance or an object of a class. In Python, it is explicitly used as the first parameter, unlike in Java where it is optional. It helps in distinguishing between the methods and attributes of a class from its local variables.

70.What is PYTHONPATH in Python?

PYTHONPATH is an environment variable which you can set to add additional directories where Python will look for modules and packages. This is especially useful in maintaining Python libraries that you do not wish to install in the global default location.

71.What is the difference between .py and .pyc files?

..

.py files contain the source code of a program. Whereas, .pyc file contains the bytecode

of your program. We get bytecode after compilation of .py file (source code). .pyc files are not created for all the files that you run. It is only created for the files that you import.

72. Explain how you can access a module written in Python from C?

You can access a module written in Python from C by following method,
Module = =PyImport_ImportModule("<modulename>");

73.What is namespace in Python?

In Python, every name introduced has a place where it lives and can be hooked for. This is known as namespace. It is like a box where a variable name is mapped to the object placed. Whenever the variable is searched out, this box will be searched, to get the corresponding object.

74.What is pickling and unpickling?

Pickle module accepts any Python object and converts it into a string representation and dumps it into a file by using the dump function, this process is called pickling. While the process of retrieving original Python objects from the stored string representation is called unpickling.

75. How is Python interpreted?

Python language is an interpreted language. The Python program runs directly from the source code. It converts the source code that is written by the programmer into an intermediate language, which is again translated into machine language that has to be executed.

Jupyter Notebook

76.What is the main use of a Jupyter notebook?

Jupyter Notebook is an open-source web application that allows us to create and share codes and documents. It provides an environment, where you can document your code, run it, look at the outcome, visualize data and see the results without leaving the environment.

77. How do I increase the cell width of the Jupyter/ipython notebook in my browser?

```
>>> from IPython.core.display import display, HTML
>>> display(HTML("<style>.container { width:100% !important; }</style>"))
```

..

78. How do I convert an IPython Notebook into a Python file via command line?

```
>>> jupyter nbconvert --to script [YOUR_NOTEBOOK].ipynb
```

79. How to measure execution time in a jupyter notebook?

```
>>> %%time is inbuilt magic command
```

80. How to run a jupyter notebook from the command line?

```
>>> jupyter nbconvert --to python nb.ipynb
```

81. How to make inline plots larger in jupyter notebooks?

Use figure size.

```
>>> fig=plt.figure(figsize=(18, 16), dpi= 80, facecolor='w', edgecolor='k')
```

82. How to display multiple images in a jupyter notebook?

```
>>>for ima in images:
>>>plt.figure()
>>>plt.imshow(ima)
```

83.Why is the Jupyter notebook interactive code and data exploration friendly?

The ipywidgets package provides many common user interface controls for exploring code

and data interactively.

84.What is the default formatting option in jupyter notebook?

Default formatting option is markdown

85.What are kernel wrappers in jupyter?

Jupyter brings a lightweight interface for kernel languages that can be wrapped in Python. Wrapper kernels can implement optional methods, notably for code completion and code inspection.

86.What are the advantages of custom magic commands?

Create IPython extensions with custom magic commands to make interactive computing even easier. Many third-party extensions and magic commands exist, for example, the %%cython magic that allows one to write Cython code directly in a notebook.

87. Is the jupyter architecture language dependent?

No. It is language independent.

..

88.Which tools allow jupyter notebooks to easily convert to pdf and html?

Nbconvert converts it to pdf and html while Nbviewer renders the notebooks on the web platforms.

89.What is a major disadvantage of a Jupyter notebook?

It is very hard to run long asynchronous tasks. Less Secure.

90. In which domain is the jupyter notebook widely used?

It is mainly used for data analysis and machine learning related tasks.

91.What are alternatives to jupyter notebook?

PyCharm interact, VS Code Python Interactive etc.

92.Where can you make configuration changes to the jupyter notebook?

In the config file located at ~/.ipython/profile_default/ipython_config.py

93.Which magic command is used to run python code from jupyter notebook?

%run can execute python code from .py files

94. How to pass variables across the notebooks?

The %store command lets you pass variables between two different notebooks.

```
>>> data = 'this is the string I want to pass to different notebook'
>>> %store data
# Stored 'data' (str)
# In new notebook
>>> %store -r data
>>> print(data)
```

95. Export the contents of a cell/Show the contents of an external script

Using the %%writefile magic saves the contents of that cell to an external file. %pycat does the opposite and shows you (in a popup) the syntax highlighted contents of an external file.

96.What inbuilt tool we use for debugging python code in a jupyter notebook?

Jupyter has its own interface for The Python Debugger (pdb). This makes it possible to go inside the function and investigate what happens there.

..

97. How to make high resolution plots in a jupyter notebook?

```
>>> %config InlineBackend.figure_format ='retina'
```

98. How can one use latex in a jupyter notebook?

When you write LaTeX in a Markdown cell, it will be rendered as a formula using MathJax.
99. What is a jupyter lab?
It is a next generation user interface for conventional jupyter notebooks. Users can drag and drop cells, arrange code workspace and live previews. It's still in the early stage of development.
100. What is the biggest limitation for a Jupyter notebook?
Code versioning, management and debugging is not scalable in current jupyter notebook.
..

Top 100 Machine Learning Questions & Answers

..

Q1 Explain the difference between supervised and unsupervised machine learning?

In supervised machine learning algorithms, we have to provide labeled data, for example, prediction of stock market prices, whereas in unsupervised we need not have labeled data, for example, classification of emails into spam and non-spam.

Q2 What are the parametric models? Give an example.

Parametric models are those with a finite number of parameters. To predict new data, you only need to know the parameters of the model. Examples include linear regression, logistic regression, and linear SVMs.

Non-parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include decision trees, k-nearest neighbors, and topic models using latent Dirichlet analysis.

Q3 What is the difference between classification and regression?

Classification is used to produce discrete results, classification is used to classify data into some specific categories. For example, classifying emails into spam and non-spam categories.

Whereas, We use regression analysis when we are dealing with continuous data, for example predicting stock prices at a certain point in time.

Q4 What Is Overfitting, and How Can You Avoid It?

Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.

When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is
known as overfitting.
There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used
- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

Q5 What is meant by 'Training set' and 'Test Set'?

We split the given data set into two different sections namely,'Training set' and 'Test Set'.
'Training set' is the portion of the dataset used to train the model.
'Testing set' is the portion of the dataset used to test the trained model.

Q6 How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or
replace them entirely with some other value.
There are two useful methods in Pandas:

- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

Q7 Explain Ensemble learning.

In ensemble learning, many base models like classifiers and regressors are generated and combined together so that they give better results. It is used when we build component classifiers that are accurate and independent. There are sequential as well as parallel ensemble
methods.

Q8 Explain the Bias-Variance Tradeoff.

Predictive models have a tradeoff between bias (how well the model fits the data) and variance
(how much the model changes based on changes in the inputs).
Simpler models are stable (low variance) but they don't get close to the truth (high bias).
More complex models are more prone to overfitting (high variance) but they are expressive

enough to get close to the truth (low bias). The best model for a given problem usually lies somewhere in the middle.

Q9 What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments.

In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution.

In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

Q10 How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit.

..

For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

Q11 What are 3 data preprocessing techniques to handle outliers?

1. Winsorize (cap at threshold).
2. Transform to reduce skew (using Box-Cox or similar).
3. Remove outliers if you're certain they are anomalies or measurement errors.

Q12 How much data should you allocate for your training, validation, and test sets?

You have to find a balance, and there's no right answer for every problem.

If your test set is too small, you'll have an unreliable estimation of model performance (performance statistic will have high variance). If your training set is too small, your actual model parameters will have a high variance.

A good rule of thumb is to use an 80/20 train/test split. Then, your train set can be further split into train/validation or into partitions for cross-validation.

Q13 What Is a False Positive and False Negative and How Are They Significant?

False positives are those cases which wrongly get classified as True but are False.

False negatives are those cases which wrongly get classified as False but are True.

In the term ‘False Positive,’ the word ‘Positive’ refers to the ‘Yes’ row of the predicted value in
the confusion matrix. The complete term indicates that the system has predicted it as a positive,
but the actual value is negative.

Q14 Explain the difference between L1 and L2 regularization.

L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with
many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a
Laplacean prior to the terms, while L2 corresponds to a Gaussian prior.

Q15 What’s a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of
symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it’s how we find
the recipe. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to
match any time signal. A Fourier transform converts a signal from time to frequency domain —
it’s a very common way to extract features from audio signals or other time series such as
sensor data.

Q16 What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to
use backpropagation and certain principles from neuroscience to more accurately model large
sets of unlabelled or semi-structured data. In that sense, deep learning represents an
..
unsupervised learning algorithm that learns representations of data through the use of neural
nets.

Q17 What’s the difference between a generative and discriminative model?

A generative model will learn categories of data while a discriminative model will simply learn
the distinction between different categories of data. Discriminative models will generally
outperform generative models on classification tasks.

Q18 What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:

● Email Spam Detection

Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.

● Healthcare Diagnosis

By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.

● Sentiment Analysis

This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral, or negative in sentiment.

● Fraud Detection

Training the model to identify suspicious patterns, we can detect instances of possible fraud.

Q19 What Is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses
no training data.
In the case of semi-supervised learning, the training data contains a small amount of labeled
data and a large amount of unlabeled data.

Q20. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

Clustering

● Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

Association

● In an association problem, we identify patterns of associations between different variables or items.

● For example, an eCommerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

..

Q21 What Is 'naive' in the Naive Bayes Classifier?

The classifier is called ‘naive’ because it makes assumptions that may or may not turn out to be
correct.
The algorithm assumes that the presence of one feature of a class is not related to the presence
of any other feature (absolute independence of features), given the class variable.
For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

Q22 Explain Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation (LDA) is a common method of topic modeling, or classifying documents by subject matter.
LDA is a generative model that represents documents as a mixture of topics that each have their own probability distribution of possible words.
The "Dirichlet" distribution is simply a distribution of distributions. In LDA, documents are distributions of topics that are distributions of words.

Q23 Explain Principle Component Analysis (PCA).

PCA is a method for transforming features in a dataset by combining them into uncorrelated linear combinations.
These new features, or principal components, sequentially maximize the variance represented
(i.e. the first principal component has the most variance, the second principal component has
the second most, and so on).
As a result, PCA is useful for dimensionality reduction because you can set an arbitrary variance cutoff.

Q24 What’s the F1 score? How would you use it?

The F1 score is a measure of a model’s performance. It is a weighted average of the precision
and recall of a model, with results tending to 1 being the best, and those tending to 0 being the
worst. You would use it in classification tests where true negatives don’t matter much.

..

Q25 When should you use classification over regression?

Classification produces discrete values and dataset to strict categories, while regression gives

you continuous results that allow you to better distinguish differences between individual points.

You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

Q26 How do you ensure you're not overfitting with a model?

This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

1- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.

2- Use cross-validation techniques such as k-folds cross-validation.

3- Use regularization techniques such as LASSO that penalize certain model parameters if they're likely to cause overfitting.

Q27 How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

Q28 How Do You Design an Email Spam Filter?

Building a spam filter involves the following process:

- The email spam filter will be fed with thousands of emails
- Each of these emails already has a label: 'spam' or 'not spam.'
- The supervised machine learning algorithm will then determine which type of emails are being marked as spam based on spam words like the lottery, free offer, no money, full refund, etc.
- The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like Decision Trees and SVM to determine how likely the email

is spam

● If the likelihood is high, it will label it as spam, and the email won't hit your inbox

..

● Based on the accuracy of each model, we will use the algorithm with the highest accuracy after testing all the models

Q29 What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the
data. You should then implement a choice selection of performance metrics: here is a fairly comprehensive list. You could use measures such as the F1 score, the accuracy, and the confusion matrix. What's important here is to demonstrate that you understand the nuances of
how a model is measured and how to choose the right performance measures for the right situations.

Q30 How would you implement a recommendation system for our company's users?

A lot of machine learning interview questions of this type will involve the implementation of machine learning models to a company's problems. You'll have to research the company and its
industry in-depth, especially the revenue drivers the company has, and the types of users the
company takes on in the context of the industry it's in.

Q31 Explain bagging.

Bagging, or Bootstrap Aggregating, is an ensemble method in which the dataset is first divided
into multiple subsets through resampling.
Then, each subset is used to train a model, and the final predictions are made through voting or
averaging the component models.
Bagging is performed in parallel.

Q32 What is the ROC Curve and what is AUC (a.k.a. AUROC)?

The ROC (receiver operating characteristic) the performance plot for binary classifiers of True
Positive Rate (y-axis) vs. False Positive Rate (xaxis).

AUC is the area under the ROC curve, and it's a common performance metric for evaluating binary classification models.
It's equivalent to the expected probability that a uniformly drawn random positive is ranked before a uniformly drawn random negative.

..

Q33 Why is Area Under ROC Curve (AUROC) better than raw accuracy as an out-of-sample evaluation metric?

AUROC is robust to class imbalance, unlike raw accuracy.
For example, if you want to detect a type of cancer that's prevalent in only 1% of the population,
you can build a model that achieves 99% accuracy by simply classifying everyone has cancer-free.

Q34 What are the advantages and disadvantages of neural networks?

Advantages: Neural networks (specifically deep NNs) have led to performance breakthroughs for unstructured datasets such as images, audio, and video. Their incredible flexibility allows
them to learn patterns that no other ML algorithm can learn.
Disadvantages: However, they require a large amount of training data to converge. It's also difficult to pick the right architecture, and the internal "hidden" layers are incomprehensible.

Q35 Define Precision and Recall.

Precision

● Precision is the ratio of several events you can correctly recall to the total number of events you recall (mix of correct and wrong recalls).
● Precision = (True Positive) / (True Positive + False Positive)

Recall

● A recall is the ratio of a number of events you can recall the number of total events.
● Recall = (True Positive) / (True Positive + False Negative)

Q36 What Is Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like
way with branches and nodes. Decision trees can handle both categorical and numerical data.

Q37 What Is Pruning in Decision Trees, and How Is It Done?

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the

complexity of the final classifier, and hence improves predictive accuracy by the reduction of
overfitting.

Pruning can occur in:

- Top-down fashion. It will traverse nodes and trim subtrees starting at the root
- Bottom-up fashion. It will begin at the leaf nodes

There is a popular pruning algorithm called reduced error pruning, in which:

- Starting at the leaves, each node is replaced with its most popular class
- If the prediction accuracy is not affected, the change is kept
- There is an advantage of simplicity and speed

..

Q38 What Is a Recommendation System?

Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system:

It's an information filtering system that predicts what a user might want to hear or see based on
choice patterns provided by the user.

Q39 What Is Kernel SVM?

Kernel SVM is the abbreviated version of the kernel support vector machine. Kernel methods
are a class of algorithms for pattern analysis, and the most common one is the kernel SVM.

Q40 What Are Some Methods of Reducing Dimensionality?

You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

Now that you have gone through these machine learning interview questions, you must have got an idea of your strengths and weaknesses in this domain.

Q41 What Are the Three Stages of Building a Model in Machine Learning?

The three stages of building a machine learning model are:

- **Model Building** Choose a suitable algorithm for the model and train it according to the requirement
- **Model Testing** Check the accuracy of the model through the test data
- **Applying the Mode** Make the required changes after testing and use the final model for real-time projects. Here, it's important to remember that once in a while, the model needs to be checked to make sure it's working correctly. It should be modified to make sure that it is up-to-date.

Q42 How is KNN different from k-means clustering?

K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this
really means is that in order for K-Nearest Neighbors to work, you need labeled data you want
to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires
only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between
different points.

Q43 Mention the difference between Data Mining and Machine learning?

Machine learning relates to the study, design, and development of the algorithms that give computers the capability to learn without being explicitly programmed. While data mining can
be defined as the process in which the unstructured data tries to extract knowledge or unknown
interesting patterns. During this processing machine, learning algorithms are used.

..

Q44 What are the different Algorithm techniques in Machine Learning?

The different types of techniques in Machine Learning are

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning
- Transduction
- Learning to Learn

Q45 You are given a data set. The data set has missing values that spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

This question has enough hints for you to start thinking! Since the data is spread across the median, let's assume it's a normal distribution. We know, in a normal distribution, ~68% of the
data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data
unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Q46 What are PCA, KPCA, and ICA used for?

PCA (Principal Components Analysis), KPCA (Kernel-based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used
for dimensionality reduction.

Q47 What are support vector machines?

Support vector machines are supervised learning algorithms used for classification and regression analysis.

Q48 What is batch statistical learning?

Statistical learning techniques allow learning a function or predictor from a set of observed data
that can make predictions about unseen or future data. These techniques provide guarantees
on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

Q49 What is the bias-variance decomposition of classification error in the ensemble method?

The expected error of a learning algorithm can be decomposed into bias and variance. A bias
term measures how closely the average classifier produced by the learning algorithm matches
the target function. The variance term measures how much the learning algorithm's prediction
fluctuates for different training sets.

..

Q50 When is Ridge regression favorable over Lasso regression?

You can quote ISLR's authors Hastie, Tibshirani who asserted that, in the presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables
with small/medium-sized effects, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all
the coefficients in the model. In the presence of correlated variables, ridge regression might be
the preferred choice. Also, ridge regression works best in situations where the least square

estimates have higher variance. Therefore, it depends on our model objective.

Q51 You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

The model has overfitted. Training error 0.00 means the classifier has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on an unseen sample, it couldn't find those patterns and returned predictions
with higher error. In a random forest, it happens when we use a larger number of trees than necessary. Hence, to avoid this situation, we should tune the number of trees using cross-validation.

Q50 What is a convex hull?

In the case of linearly separable data, the convex hull represents the outer boundaries of the
two groups of data points. Once the convex hull is created, we get maximum margin hyperplane
(MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to
create the greatest separation between two groups.

Q51 What do you understand by Type I vs Type II error?

Type I error is committed when the null hypothesis is true and we reject it, also known as a 'False Positive'. Type II error is committed when the null hypothesis is false and we accept it,
also known as 'False Negative'.
In the context of the confusion matrix, we can say Type I error occurs when we classify a value
as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as
negative (0) when it is actually positive(1).

Q52. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance?

We don't use manhattan distance because it calculates distance horizontally or vertically only. It
has dimension restrictions. On the other hand, the euclidean metric can be used in any space to

calculate distance. Since the data points can be present in any dimension, euclidean distance is
a more viable option.

..

Example: Think of a chessboard, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

Q53 Do you suggest that treating a categorical variable as a continuous variable would result in a better predictive model?

For better predictions, the categorical variable can be considered as a continuous variable only
when the variable is ordinal in nature.

Q54 OLS is to linear regression. The maximum likelihood is logistic regression. Explain the statement.

OLS and Maximum likelihood are the methods used by the respective regression methods to
approximate the unknown parameter (coefficient) value. In simple words,
Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the values of parameters which maximizes the likelihood that the
parameters are most likely to produce observed data.

Q55 When does regularization becomes necessary in Machine Learning?

Regularization becomes necessary when the model begins to overfit/underfit. This technique
introduces a cost term for bringing in more features with the objective function. Hence, it tries to
push the coefficients for many variables to zero and hence reduce the cost term. This helps to
reduce model complexity so that the model can become better at predicting (generalizing).

Q56 What is Linear Regression?

Linear Regression is a supervised Machine Learning algorithm. It is used to find the linear relationship between the dependent and the independent variables for predictive analysis.

Q57 What is the Variance Inflation Factor?

Variance Inflation Factor (VIF) is the estimate of the volume of multicollinearity in a collection of
many regression variables.

VIF = Variance of the model / Variance of the model with a single independent variable
We have to calculate this ratio for every independent variable. If VIF is high, then it shows the
high collinearity of the independent variables.

Q58 We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

When we use one-hot encoding, there is an increase in the dimensionality of a dataset. The reason for the increase in dimensionality is that, for every class in the categorical variables, it
forms a different variable.

..

Q59 What is a Decision Tree?

A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.

Q60 What is the Binarizing of data? How to Binarize?

In most of the Machine Learning Interviews, apart from theoretical questions, interviewers focus
on the implementation part. So, this ML Interview Questions focused on the implementation of
the theoretical concepts.

Converting data into binary values on the basis of threshold values is known as the binarizing of
data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when we have to perform feature engineering, and we can also use it for adding unique features.

Q61 What is cross-validation?

Cross-validation is essentially a technique used to assess how well a model performs on a new
independent dataset. The simplest example of cross-validation is when you split your data into
two groups: training data and testing data, where you use the training data to build the model
and the testing data to test the model.

Q62 When would you use random forests Vs SVM and why?

There are a couple of reasons why a random forest is a better choice of the model than a support vector machine:

● Random forests allow you to determine the feature importance. SVM's can't do this.
● Random forests are much quicker and simpler to build than an SVM.
● For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

Q63 What are the drawbacks of a linear model?

There are a couple of drawbacks of a linear model:
● A linear model holds some strong assumptions that may not be true in the application. It assumes a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity
● A linear model can't be used for discrete or binary outcomes.
● You can't vary the model flexibility of a linear model.

..

Q64 Do you think 50 small decision trees are better than a large one? Why?

Another way of asking this question is "Is a random forest a better model than a decision tree?"
And the answer is yes because a random forest is an ensemble method that takes many weak
decision trees to make a strong learner. Random forests are more accurate, more robust, and
less prone to overfitting.

Q65 What is a kernel? Explain the kernel trick

A kernel is a way of computing the dot product of two vectors x and y in some (possibly very
high dimensional) feature space, which is why kernel functions are sometimes called "generalized dot product"
The kernel trick is a method of using a linear classifier to solve a non-linear problem by transforming linearly inseparable data to linearly separable ones in a higher dimension.

Q66 State the differences between causality and correlation?

Causality applies to situations where one action, say X, causes an outcome, say Y, whereas Correlation is just relating one action (X) to another action(Y) but X does not necessarily cause
Y.

Q67 What is the exploding gradient problem while using the backpropagation technique?

When large error gradients accumulate and result in large changes in the neural network

weights during training, it is called the exploding gradient problem. The values of weights can
become so large as to overflow and result in NaN values. This makes the model unstable and
the learning of the model to stall just like the vanishing gradient problem.

Q68 What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated.

Q69 What is Marginalisation? Explain the process.

Marginalizationarginalisation is summing the probability of a random variable X given the joint
probability distribution of X with other variables. It is an application of the law of total probability.

Q70 Why is the rotation of components so important in Principle Component Analysis(PCA)?

Rotation in PCA is very important as it maximizes the separation within the variance obtained by
all the components because of which interpretation of components would become easier. If the
components are not rotated, then we need extended components to describe the variance of
the components.

..

Q71 What is the difference between regularization and normalisation?

Normalisation adjusts the data; regularisation adjusts the prediction function. If your data is on
very different scales (especially low to high), you would want to normalise the data. Alter each
column to have compatible basic statistics. This can be helpful to make sure there is no loss of
accuracy. One of the goals of model training is to identify the signal and ignore the noise if the
model is given free rein to minimize error, there is a possibility of suffering from overfitting.
Regularization imposes some control on this by providing simpler fitting functions over complex
ones.

Q72 When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

A place where the highest RSquared value is found, is the place where the line comes to rest.

RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

Q73 How does the SVM algorithm deal with self-learning?

SVM has a learning rate and expansion rate which takes care of this. The learning rate compensates or penalises the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

Q74 How do you handle outliers in the data?

Outlier is an observation in the data set that is far away from other observations in the data set.

We can discover outliers using tools and functions like box plot, scatter plot, Z-Score, IQR score etc. and then handle them based on the visualization we have got. To handle outliers, we can cap at some threshold, use transformations to reduce skewness of the data and remove outliers if they are anomalies or errors.

Q75 Name and define techniques used to find similarities in the recommendation system.

Pearson correlation and Cosine correlation are techniques used to find similarities in recommendation systems.

Q76 Why would you Prune your tree?

In the context of data science or AIML, pruning refers to the process of reducing redundant branches of a decision tree. Decision Trees are prone to overfitting, pruning the tree helps to reduce the size and minimizes the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the tradeoff.

..

Q77 Mention some of the EDA Techniques?

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the

foundation of better models.

Visualization

- Univariate visualization
- Bivariate visualization
- Multivariate visualization

Missing Value Treatment – Replace missing values with Either Mean/Median

Outlier Detection – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the
boundary for IQR

Q78 What is data augmentation? Can you give some examples?

Data augmentation is a technique for synthesizing new data by modifying existing data in such a
way that the target is not changed, or it is changed in a known way.

CV is one of the fields where data augmentation is very useful. There are many modifications
that we can do to images:

- Resize
- Horizontal or vertical flip
- Rotate
- Add noise
- Deform
- Modify colors

Each problem needs a customized data augmentation pipeline. For example, on OCR, doing flips will change the text and won't be beneficial; however, resizes and small rotations may help.

Q79 What is Inductive Logic Programming in Machine Learning (ILP)?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logic programming representing background knowledge and examples.

Q80 What is the difference between inductive machine learning and deductive machine learning?

The difference between inductive machine learning and deductive machine learning are as follows: machine-learning where the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws
the conclusion and then the conclusion is drawn.

..

Q81 Difference between machine learning and deep learning

Machine learning is a branch of computer science and a method to implement artificial intelligence. This technique provides the ability to automatically learn and improve from experiences without being explicitly programmed.

Deep learning can be said as a subset of machine learning. It is mainly based on the artificial neural network where data is taken as an input and the technique makes intuitive decisions using the artificial neural network.

Q82 What Are The Steps Involved In Machine Learning Project?

As you plan for doing a machine learning project. There are several important steps you must follow to achieve a good working model and they are data collection, data preparation, choosing a machine learning model, training the model, model evaluation, parameter tuning and lastly prediction.

Q83 Differences between Artificial Intelligence and Machine Learning?

Artificial intelligence is a broader prospect than machine learning. Artificial intelligence mimics the cognitive functions of the human brain. The purpose of AI is to carry out a task in an intelligent manner based on algorithms. On the other hand, machine learning is a subclass of artificial intelligence. To develop an autonomous machine in such a way so that it can learn without being explicitly programmed is the goal of machine learning.

Q84 Steps Needed to Choose the Appropriate Machine Learning Algorithm for your Classification problem.

Firstly, you need to have a clear picture of your data, your constraints, and your problems before heading towards different machine learning algorithms. Secondly, you have to understand which type and kind of data you have because it plays a primary role in deciding which algorithm you have to use.

Following this step is the data categorization step, which is a two-step process – categorization by input and categorization by output. The next step is to understand your constraints; that is, what is your data storage capacity? How fast the prediction has to be? etc.

Finally, find the available machine learning algorithms and implement them wisely. Along with

that, also try to optimize the hyperparameters which can be done in three ways – grid search,
random search, and Bayesian optimization.

..

Q85 Explain Backpropagation in Machine Learning.

A very important question for your machine learning interview. Backpropagation is the algorithm for computing artificial neural networks (ANN). It is used by the gradient descent optimization that exploits the chain rule. By calculating the gradient of the loss function, the weight of the neurons is adjusted to a certain value. To train a multi-layered neural network is
the prime motivation of backpropagation so that it can learn the appropriate internal demonstrations. This will help them learn to map any input to its respective output arbitrarily.

Q86 What is the Convex Function?

This question is very often asked in machine learning interviews. A convex function is a continuous function, and the value of the midpoint at every interval in its given domain is less
than the numerical mean of the values at the two ends of the interval.

Q87 What's the Relationship between True Positive Rate and Recall?

The True positive rate in machine learning is the percentage of the positives that have been properly acknowledged, and recall is just the count of the results that have been correctly identified and are relevant. Therefore, they are the same things, just having different names. It is
also known as sensitivity.

Q88 List some Tools for Parallelizing Machine Learning Algorithms.

Although this question may seem very easy, make sure not to skip this one because it is also
very closely related to artificial intelligence and thereby, AI interview questions. Almost all machine learning algorithms are easy to serialize. Some of the basic tools for parallelizing are
Matlab, Weka, R, Octave, or the Python-based sci-kit learn.

Q89 What do you mean by Genetic Programming?

Genetic Programming (GP) is almost similar to an Evolutionary Algorithm, a subset of machine
learning. Genetic programming software systems implement an algorithm that uses random mutation, a fitness function, crossover, and multiple generations of evolution to resolve a

user-defined task. The genetic programming model is based on testing and choosing the best
option among a set of results.

Q90 What do you know about Bayesian Networks?

Bayesian Networks also referred to as 'belief networks' or 'casual networks', are used to represent the graphical model for probability relationship among a set of variables.

For example, a Bayesian network can be used to represent the probabilistic relationships between diseases and symptoms. As per the symptoms, the network can also compute the probabilities of the presence of various diseases.

..

Efficient algorithms can perform inference or learning in Bayesian networks. Bayesian networks
which relate the variables (e.g., speech signals or protein sequences) are called dynamic Bayesian networks.

Q91 Which are the two components of the Bayesian logic program?

A Bayesian logic program consists of two components:

● **Logical** It contains a set of Bayesian Clauses, which capture the qualitative structure of the domain.

● **Quantitative** It is used to encode quantitative information about the domain.

Q92 How is machine learning used in day-to-day life?

Most of the people are already using machine learning in their everyday life. Assume that you
are engaging with the internet, you are actually expressing your preferences, likes, dislikes through your searches. All these things are picked up by cookies coming on your computer, from this, the behavior of a user is evaluated. It helps to increase the progress of a user through
the internet and provide similar suggestions.

The navigation system can also be considered as one of the examples where we are using machine learning to calculate a distance between two places using optimization techniques. Surely, people are going to more engage with machine learning in the near future

Q93 Define Sampling. Why do we need it?

Answer: Sampling is a process of choosing a subset from a target population that would serve
as its representative. We use the data from the sample to understand the pattern in the community as a whole. Sampling is necessary because often, we can not gather or process the

complete data within a reasonable time.

Q94 What does the term decision boundary mean?

Answer: A decision boundary or a decision surface is a hypersurface which divides the underlying feature space into two subspaces, one for each class. If the decision boundary is a
hyperplane, then the classes are linearly separable.

Q95 Define entropy?

Answer: Entropy is the measure of uncertainty associated with random variable Y. It is the expected number of bits required to communicate the value of the variable.

Q96 Indicate the top intents of machine learning?

Answer: The top intents of machine learning are stated below,

● The system gets information from the already established computations to give well-founded decisions and outputs.

● It locates certain patterns in the data and then makes certain predictions on it to provide answers on matters.

..

Q97 Highlight the differences between the Generative model and the Discriminative model?

The aim of the Generative model is to generate new samples from the same distribution and new data instances, Whereas, the Discriminative model highlights the differences between different kinds of data instances. It tries to learn directly from the data and then classifies the
data.

Q98 Identify the most important aptitudes of a machine learning engineer?

Machine learning allows the computer to learn itself without being decidedly programmed. It helps the system to learn from experience and then improve from its mistakes. The intelligence
system, which is based on machine learning, can learn from recorded data and past incidents.

In-depth knowledge of statistics, probability, data modelling, programming language, as well as
CS, Application of ML Libraries and algorithms, and software design is required to become a successful machine learning engineer.

Q99 What is feature engineering? How do you apply it in the process of modelling?

Feature engineering is the process of transforming raw data into features that better represent
the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Q100 How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.
In a learning curve, the training error and cross-validating error are plotted against the number
of training data points.

The inception network contains nine inception blocks. These nine inception blocks are stacked one above the other. First, we take the input image and we perform the convolutional operation with three filters of varying sizes which include 1 x 1, 3 x3, and 5 x5. Then we feed the result of the convolutional operation to the next inception block.

Why 1 x 1 convolution is useful?

1x1 convolution implies that we use one filter of size 1 x 1. It is widely used for reducing the number of depth channels.

What is called factorized convolution?

We can break down a convolutional layer with a larger filter size into a stack of convolutional layers with smaller filter size and this is known as factorized convolution.
Suppose, we have a convolutional layer with a 5 x 5 filter then it can be broken down into two convolutional layers with 3 x 3 filters.

Explain the architecture of LeNet

The LeNet architecture consists of seven layers as given below:

a)Three convolutional layers
b)Two pooling layers
c)One fully connected layer
d)One output layer

What are the drawbacks of CNN?

CNN is translation-invariant and this makes CNN more prone to misclassification. Say, for instance, we are performing a face recognition task then CNN checks only the presence of facial features such as eyes, nose, mouth, and ears. It will not check whether those features are present in the correct locations. If the images have all those features, then it will be classified as the face irrespective of the location of the features. This is one of the major drawbacks of CNN.

Why CNN is most preferred for the image data?

The convolutional neural network uses the special operation called convolution which is capable of extracting important features from the image. Since the convolutional operation extracts good features from the image, the accuracy of CNN is high compared to the other algorithms for the image data.

ML, DL Questions

1. How you can define Machine Learning?
2. What do you understand Labelled training dataset?
3. What are 2 most common supervised ML tasks you have performed so far?
4. What kind of Machine learning algorithm would you used to walk robot in various unknown area?
5. What kind of ML algo you can use to segment your user into multiple groups?
6. What type of learning algo realised on similarity measure to make a prediction?
7. What is an online learning system?
8. What is out of core learning?
9. Can you name couple of ml challenges that you have faced?
10. Can you please give 1 example of hyperparameter tuning wrt some classification algorithm?
11. What is out of bag evaluation?
What do you understand by hard & soft voting classifier?
13. Let's suppose your ML algorithm is taking 5 min time to train, How will you bring down time to 5 second for
training? (Hint: Distributed Computation)
14. Let's Suppose I have trained 5 diff model with same training dataset & all of them have achieved 95%
precision. Is there any chance that you can combine all these models to get better result? If yes, How? If no,
Why?
15. What do you understand by Gradient decent? How will you explain Gradient decent to a kid?
Can you please explain diff between regression & classification?
Explain a clustering algorithm of your choice.
How you can explain ML, DL, NLP, Computer vision & reinforcement learning with example in your own
terms?
19. How you can explain semi-supervised ML in your own way with example?
20. What is difference between abstraction & generalization in your own word.
21. What are the steps that you have followed in your last project to prepare the dataset?
In your last project what steps were involved in model selection procedure?
23. If I give you 2 columns of any dataset, what will be the steps will be involved to check the relationship
between those 2 columns?
24. Can you please explain 5 diff kind of strategies at least to handle missing values in dataset?
25. What kind of diff. issues you have faced wrt your raw data? At least mention 5 issues.
What is your strategy to handle categorical dataset? Explain with example.
27. How do you define a model in terms of machine learning or in your own word?
28. What do you understand by k fold validation & in what situation you have used k fold cross validation?
What is meaning of bootstrap sampling? explain me in your own word.
What do you understand by underfitting & overfitting of model with example?
What is diff between cross validation and bootstrapping?
What do you understand by silhouette coefficient?

What is the advantage of using ROC Score?
34. Explain me complete approach to evaluate your regression model
Give me example of lazy learner and eagar learner algorithms example.
What do you understand by holdout method?
What is diff between predictive modelling and descriptive modelling.
How you have derived a feature for model building in your last project?
Explain 5 different encoding techniques.
How do you define some features are not important for ML model? What strategy will you follow
What is difference between Euclidian distance and Manhattan distance. Explain in simple words.
What do you understand by feature selection, transformation, engineering and EDA & What are the steps
that you have performed in each of these in detail with example.
What is difference between single values decomposition (SVD) and PCA? (hint: SVD is one of the way to do
PCA)
What kind of feature transformations you have done in your last project?
45. Have you taken any external feature in any of project from any 3rd party data? If yes, explain that scenario.
If your model is overfitted, what you will do next?
Explain me bias variance trade-off.
What steps would you take to improve accuracy of your model? At-least mention 5 approach. And justify
why would you choose those approach
Explain process of feature engineering in context of text categorization.
Explain vectorization and hamming distance.
Can you please explain chain rule and its use?
What is difference between correlation and covariance?
What are the sampling techniques you have used in your project?
Have you ever used Hypothesis testing in your last project, if yes, explain How?
In which case you will use naïve Bayes classifier and decision tree separately?
What is the adv & disadvantage of naïve Bayes classifier, explain
In case of numerical data what is naïve Bayes classification equation you will use?
Give me scenario where I will be able to use a boosting classifier and regressor?
In case of Bayesian classifier what exactly it tries to learn. Define its learning procedure.
Give me a situation where I will be able to use SVM instead of Logistic regression.
What do you understand by rbf kernel in SVM?
Give me 2 scenarios where AI can be used to increase revenue of travel industry.
What do you understand by leaf node in decision tree?
What is information gain & Entropy in decision tree?
Give disadvantages of using Decision tree
List some of the features of random forest.
How can you avoid overfitting in decision tree?
Explain polynomial regression in your own way.
Explain learning mechanism of linear regression.
What is the cost function in logistic regression?
What is the error function in linear regression?
What is the use of implementing OLS technique wrt dataset?
Explain dendrogram in your own way.
How do you measure quality of clusters in DBSCAN?
How do you evaluate DBSCAN algorithm?
What do you understand by market basket analysis?
Explain centroid formation technique in K Means algorithm.
Have you ever used SVM regression in any of your project, If yes, Why?

Explain the concept of GINI Impurity.
Let’s suppose I have given you dataset with 100 column how you will be able to control growth of decision
tree?
If you are using Ada-boost algorithm & if it is giving you underfitted result What is the hyperparameter tuning
you will do?
Explain gradient boosting algorithm.
Can we use PCA to reduce dimensionality of highly non-linear data.
How do you evaluate performance of PCA.
Have you ever used multiple dimensionality techniques in any project? if yes, give reason. If no, where can
we use it?
What do you understand by curse of dimensionality explain with help of example
What is the difference between anomaly detection and novelty detection
Explain gaussian mixture model.
Give me list of 10 activation functions with explanation
Explain neural network in terms of mathematical function
Can you please corelate a biological neuron and artificial neuron?
Give list of cost functions you heard of, with explanation.
Can I solve problem of classification with tabular data in neural network?
What do you understand by backword propagation in neural network?
Why do we need neural network instead of straight forward mathematical equation?
What are the different weight initialization techniques you have used?
Can you visualize a neural network? if yes provide name of software we can use?
How will you explain training of neural network?
Can you please explain difference between sigmoid & tanh function.
100. Explain disadvantage of using RELU function.
101. How do you select no. of layers & no. of neurons in neural network?
102. Have you ever designed any Neural network architecture by yourself?
103. Can you please explain SWISS Function?
104. What is learning rate in laymen way and how do you control learning rate?
105. What is diff between batch, minibatch & stochastic gradient decent.
106. What do you understand by batch size while training Neural N/w with example
107. Explain 5 best optimizer you know with mathematical explanation.
108. Can you build Neural network without using any library? If yes, prove it.
109. What is use of biases in neural network?
110. How do you do hyper-parameter tuning for neural network
111. What kind of regularization you used wrt neural network.
112. What are the libraries you have used for neural network implementation?
113. What do you understand by custom layer and a custom model?
114. How do you implement differentiation using TensorFlow or Pytorch library?
115. What is meaning of epoch in simple terms?
116. What do you understand by a TensorFlow record?
117. Explain the technique for doing data augmentation in deep learning
118. List down diff CNN network you heard of.
119. List down a names of object detection algorithm you know
120. What is difference between object detection and classification?
121. List down major tasks we perform in CNN.
122. List down algorithms for segmentation
123. Which algorithm you can use to track a football in football match.
124. If I give you a satellite image data, so which algo you will use to identify image from those image data

125. Which algorithm you will use for PCB fault detection.
126. What do you understand by pretrained model?
127. Explain different types of transfer learning.
128. Explain me where your CNN network will fail with example. And where we can use RNN network.
129. Which GPU you have been using to train your object detection model?
130. How much data set you have used for this model, what was epoch, time and accuracy of the model
131. What kind of optimization you have done for training object detection model
132. How do you evaluate your object detection model?
133. List down algorithm for object tracking
134. What do you understand by FPS (frame per second)?
135. Can you please explain 2D & 3D convolution?
136. What do you understand by batch normalization?
137. Which algorithm you use for detecting handwriting detection?
138. Explain me SoftMax function.
139. What is disadvantage of using RNN?
140. List down at least 5 RNN?
141. Explain architectural diagram of LSTM, Also list Adv & dis adv
142. Explain architectural diagram of BI LSTM, Also list Adv & dis adv
143. Explain architectural diagram of stacked LSTM. Also list Adv & dis adv
144. What do you understand by TF-IDF
145. How you will be able to create a Word 2 vector of your own
146. List down at least 5 vectorization technique.
147. What is difference between RNN and Encoder-Decoder.
148. What do you understand by attention mechanism and what is use of it
149. Have you read a research paper Attention you all need? If not, then why you are claiming you know NLP
150. What do you understand by multi headed attention? Explain

ML Questions Part 2

1. Tell me something about your project you have done in past?
2. What was your Dataset size for ML Project?
3. What is type of your dataset?
4. What was frequency of your dataset? (E.g. batch, streaming etc)
5. What was source system for your dataset? (E.g. sensor, satellite Kafka, cloud, etc.)
6. What was kind of derived dataset that you have mentioned in project?
7. How you have done validation dataset?
8. Have you created any pipeline to validated this dataset or you were using any tool?
9. What do you understand by data lake?
10. What do you understand by data warehousing?
11. Can you please name some validations that you have done on top of your data?
12. How you have handled streaming dataset?
13. How many different types of environments were available in your project?
14. What was your delivery mechanism for particular project?
15. Have you used any OPS pipeline for this current project?
16. How you were doing model retraining?
17. How you have implemented model retraining in your project?

18. How frequently you have been doing model retraining and what was the strategy for model retraining?
19. What was kind of evaluation you were doing in production environment
20. What was no. of request (hits) your model was receiving on daily basis?
21. How you have implemented logging in project for any failure cases?
22. How you have integrated notification (or Alarm) system for your project?
23. How you have implemented model monitoring?
24. How you have derived final KPI (Key Performance Indicator) for your client?
25. How many dashboards were there in your project?
26. On which platform you have productionised your model?
27. What kind of API you have exposed to receive data for model?
28. What was size of your final production environment (system configuration)?
29. What an all Databases you have used in project?
30. What kind of optimization you have done in your project, till what depth & explain the example
31. Can you please talk about complete team structure and team size?
32. What was duration of your complete project?
33. What was your day to day responsibility in last 2 month?
34. What kind of change request you have been receiving after you productionised project
35. What kind of testing you have done in development, UAT, pre-pod and prod?
36. Have you used some of the predefined AI-OPS pipelines if yes explain
37. Who has implemented AI-OPS in your project?
38. What was OPS stack you have been using?
39. What do you understand by CI-CD & have you implemented those in your project. If yes, what was the tech
stack you used for CI-CD pipeline?
40. What was biggest challenge you faced in project and how you have resolved it?
41. Give me one scenario where you worked as team player?
42. What was your overall learning from current project?
43. How do you keep yourself updated for new technology?
44. Have you designed an architecture for this project? If yes, define a strategy wrt to your current project.

-------- Below are some questions for people with 7+ years of Exp----------

What kind of discussions generally wrt client?
46. What was your contribution in team building?
How you have defined completed tech stack for AI?
What was kind of benefit you have given to your current company in terms of cost cutting and revenue?
What kind of new innovation you have introduced?
How do you push your team for research or new implementation?
How many projects you are handling?
How many clients you have acquired?
If new demand comes from client, how do you evaluate that requirement?
How do you prepare costing for project?
What kind of skillsets you look in person, to handle delivery of upcoming project?
On an average, how much time you took for building a new team?
What kind of stack you involved in initial project?
How do you decide timeline for project delivery?
How do you keep track of project progress?
How do you handle dependencies between the team?
How much profit you have given to your previous organization.

Deep learning & Vision Questions

How many images you have taken to train your DL model?
What is size of your model that you have productionised?

Have you tried optimizing this Vision or DL model?
65. Where you have hosted your Computer Vison model?
What was your frame per second?
What are the data filtration strategy you have defined for CV project in production?
Have you used any edge device in this project, if yes, why?
What was name of camera & camera quality?
What was final outcome you were generating from these devices?
Have you processed the data in local system or in cloud? Give reason.
72. How many number devices you have productionised (camera, edge devices etc.)
Let's suppose I am trying to build solution to count no. of vehicle or to detect their no. plate or track their
speed. Then what is the dependency of distance, position & angel of camera on your final model? What will
happen to your model? if we change position angle.
What was your data collection strategy in CV project, have you received data from client or you have created the data? And how you have implemented it?
What was data labelling tool that you have used for your project?
If I have to do OCR then which API you will use or you have used in your previous project?
Suppose if my images data is blur, what will be your strategy to enhance the image quality?
78. Have you implemented object tracking in any of your project, if yes, give me scenario?
Suppose there are 2 mobile devices which are moving, so suppose if this two device position overlaps with
each other then what will be your strategy to avoid any error while tracking those devices using camera?
Have you implemented multicamera tracking, do you have Any idea about it?
Explain me some of real-life use case of segmentation.
What kind of AI application you will build for retail seller to increase their sell?
Let's suppose I am trying to build AI solution to monitor a productivity of a kid, what kind of feature you
would like to give in that product?

------- NLP-------

84. Have you productionised BERT Model, if yes, can you talk about hurdles that you have faced?
How you have optimized your BERT based solution
What kind of NLP tasks you were doing wrt BERT?
Whether you have implemented BERT base & BERT large?
What are the disadvantages of using BERT?
Can please one of the lighter versions of transformer-based model?
What was the accuracy that you were receiving wrt specific tasks?
Let's suppose I have to build language-based model and there are online solution providers are there but
they are costly, so what will be your strategy?
Have you used hugging phase APIs?
What is diff between BERT and GPT based models?
There is no Decoder model in BERT then how do we get output if its just a encoder lever model?
How masking is been implemented in BERT based model and what are its disadvantages
How masking is been implemented in GPT based model and what are its disadvantages
How backpropagation happens in BERT model. Explain.
Can you explain Query, Key & Value in any Transformer based model?
What are the main reasons behind success of transformer-based models?
100. Can we use BERT based model to generate embedding? if yes, How? NO, then Why?
101. What do you think about Open AI GPT3?
102. What is your thought about convolution auto encoder?

103. List down text summarization techniques, and which latest model you will prefer for text summarization?
104. What is the meaning of multiheaded attention
105. What do you understand by BLUE Score?
106. What is gradient clipping?
107. Can you please list down ways by which I will be able to split training across multiple GPUs?
108. Explain me difference between GRU and LSTM.
109. If I have to implement Any hour then what will be your approach? (Hint: using NLTK, Spacy or state of art
model)
110. What do you understand by Uni-gram, bi-gram and tri-gram(N-gram)?
111. What do you understand by stemming and lemmatization?
112. For conversational AI solution will you use to predefined framework or will you create your own. In both
cases what are adv and disadvantages?
113. Have you worked on google dialog flow, Azure-LUIS, IBM-Watson, or RASA- NLU?
114. What are the limitations of these respective platforms mentioned above?
115. If I have to build a ticket rerouting system for a banking client, how will you design this complete system?
116. Can you please how will you be able to design an app like In-Shorts? (hint: text summarization & etc)
117. If you have to build a solution that can generate a summary of the entire online class meeting then what will
be your approach and what kind of hurdles you may face?
118. If I have to create a Gmail kind of text generation system then what will be your approach?
119. If I have to create a document parsing and validation system for legal, what will be your approach?
120. If you have to build a voice-based automation system, how you will design system architecture?

----- Time Series ----

121. List down, time series algorithms that you know?
122. How can we solve TS problem in deep learning?
123. Give application of TS in weather, financial, healthcare & network analysis?
124. What is diff between uptrend and downtrend in TS?
125. What do you understand by seasonality in
126. What do you understand by Cyclic pattern in your TS data?
127. How will you find Trend in TS Data?
128. Have you implemented ARCH model in TS? If yes, give scenario?
129. What is VAR (vector autoregressive) model?
130. What do you understand by univariant and multivariant TS Analysis?
131. Give example where you have created a multivariant model?
132. What do you understand by p, d, & q in ARIMA model
133. Tell me mechanism by which I can find p, d, q in ARIMA model?
134. What is SARIMA and how it's different from ARIMA?
135. What is meaning of AR, MA and I in ARIMA model?
136. Can we solve TS problems with transformers? what is your thought on that? why do you think in that way?
137. Have you ever productionised TS Based Model using LSTM? What are adv and disadvantages
138. Can we solve TS problem using Regressive algorithm, if yes, why, if no, give a reason?

..